DOCUMENT RESUME

ED 388 725 TM 024 203

AUTHOR Myford, Carol M.; Mislevy, Robert J.
TITLE Monitoring and Improving a Portfolio Assessment System.
INSTITUTION Educational Testing Service, Princeton, NJ. Center for Performance Assessment.; National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY College Entrance Examination Board, Princeton, N.J.; Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO ETS-MS-94-05
PUB DATE 95
CONTRACT R117G10027
NOTE 98p.
AVAILABLE FROM Center for Performance Assessment, Educational Testing Service, Mail Stop 11-P, Rosedale Road, Princeton, NJ 08541-0001.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Advanced Placement; *Art Products; *Educational Assessment; Educational Improvement; *Evaluation Methods; Interrater Reliability; Models; Naturalistic Observation; *Portfolio Assessment; Portfolios (Background Materials); Profiles; *Test Construction; Testing Programs
IDENTIFIERS Advanced Placement Examinations (CEEB); FACETS Computer Program; *Large Scale Programs; Performance Based Evaluation

ABSTRACT

Establishing and refining a framework for performance assessment is especially difficult in large-scale settings that can involve hundreds of judges and thousands of students. This presentation advocates the interactive use of two complementary analytic perspectives and illustrates the approach in the context of the College Entrance Examination Board's Advanced Placement (AP) Studio Art portfolio assessment. The "naturalistic" component of the project involved in-depth discussions with judges about 18 portfolios from the 1992 assessment that received discrepant ratings. Since it is impossible to hold such discussions for each of the individual ratings produced in the assessment, summary results for each, in the form of numerical ratings, provided the data for the "statistical" component. J. M. Linacre's (1989) FACETS model was used to summarize overall effect patterns, quantify the evidence associated with these effects, and highlight rating profiles and judge/portfolio combinations that are unusual in light of typical patterns. This focused attention where it was apt to be most useful in improving the process. By using both perspectives, one can better communicate the meaning, value, and quality of the assessment. Two appendixes contain the AP Studio Art Reader Survey and the judge interview protocol. (Contains 11 tables, 3 figures, and 26 references.) (SLD)

ED 388 725

# CENTER FOR PERFORMANCE ASSESSMENT

## Monitoring and Improving a Portfolio Assessment System

Carol M. Myford
Robert J. Mislevy

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CAROL MYFORD

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Educational
Testing Service
Princeton,
New Jersey 08541

TM024203

# Monitoring and Improving a Portfolio Assessment System

Carol M. Myford
Robert J. Mislevy

Additional copies of this report can be ordered from:

Center for Performance Assessment
Educational Testing Service
Mail Stop 11-P
Rosedale Road
Princeton, NJ 08541-0001
(609) 734-5521

MS # 94-05

Copyright © 1995 by Educational Testing Service. All rights reserved.

# Acknowledgments

This work was supported by multiple sources as follows:

- the Program Research Planning Council of Educational Testing Service,
- the College Entrance Examination Board Advanced Placement Programs,
- the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Educational Research and Development Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education, and
- the Center for Performance Assessment at Educational Testing Service.

We are grateful to J. Michael Linacre for advice and assistance on running FACETS, to David Anderson, Behroz Maneckshana, and Rick Morgan for technical assistance, and to Drew Gitomer, Rick Morgan, Howard Wainer, and Ming Mei Wang for helpful comments on an earlier draft of the paper. Alice Sims-Gunzenhauser and Ray Wilkins were instrumental throughout the project--helping us design the study, lead reader discussions, and sharpen our understanding of the rating process. We especially thank the AP Studio Art readers and Chief Reader Dan Britton, without whose cooperation and enthusiasm this project could never have succeeded.

This report is issued jointly by the Center for Performance Assessment at Educational Testing Service and the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Educational Research and Development Program.

# Abstract

An assessment system can succeed in provoking productive and sustained performances, yet fail to support instruction and evaluation unless shared standards exist among students, teachers, and judges as to what is valued in performance and how it fits into an evaluative framework. Establishing and refining such a framework is especially difficult in large-scale settings that can involve hundreds of judges and thousands of students. This presentation advocates the interactive use of two complementary analytic perspectives and illustrates the approach in the context of the College Entrance Examination Board's Advanced Placement Studio Art portfolio assessment. The "naturalistic" component of the project involved in-depth discussions with judges about 18 portfolios from the 1992 assessment that received discrepant ratings. These discussions provided insights into the kinds of evidence, inference, arguments, and standards that underlie ratings. Since it is impossible to hold such discussions for each of the 50,000+ individual ratings produced in the assessment, summary results for each, in the form of numerical ratings, provided the data for the "statistical" component. Linacre's (1989) FACETS model was used to (i) summarize overall patterns in terms of effects for students, judges, and portfolio sections, (ii) quantify the weight of evidence associated with these effects, and (iii) highlight rating profiles and judge/portfolio combinations that are unusual in light of the typical patterns. This focuses attention where it is apt to be most useful in improving the process (e.g., in clarifying expectations to students, improving judge training, or sharpening the definition of standards). Further, by making public the materials and results of both perspectives, one can better communicate the meaning and value of the work such an assessment engenders, and the quality of the processes by which evidence about students' accomplishments is evaluated.

Key words: AP Studio Art, FACETS, performance assessment, portfolio assessment, quality control, test theory, Iten Response Theory (IRT), qualitative research, arts assessment, Rasch measurement.

# Table of Contents

Acknowledgments ..... iii

Abstract ..... v

Introduction ..... 1

Background ..... 2

The AP Studio Art Program ..... 4

Discussion of Selected Portfolios ..... 15

The Statistical Analysis ..... 42

Conclusion ..... 70

References ..... 72

Appendix A ..... 74

Appendix B ..... 83

Endnotes ..... 87

# Introduction

> *As compared to measurement, assessment is inevitably involved with questions of what is of value, rather than simple correctness. Questions of value require entry and discussion. In this light, assessment is not a matter for outside experts to design; rather, it is an episode in which students and teachers might learn, through reflection and debate, about the standards of good work and the rules of evidence.*
>
> Wolf, Bixby, Glenn, & Gardner, 1991, pp. 51-52.

Performance assessment commands attention partly because it provides direct evidence about constructive and productive aspects of knowledge, and partly because it promises to improve educational practice (Resnick & Resnick, 1989). But can assessment systems founded on judgments, inevitably personal and unique, be demonstrably principled and fair nevertheless? Performance assessment challenges methodologists to devise observational situations that evoke evidence about what we want to infer, to learn how to extract and summarize this evidence, and how to monitor and improve assessment systems. This paper illustrates how the interplay of statistical and qualitative analyses can help one develop, monitor, and continually improve large-scale performance assessment systems. To this end, we analyze rating data from the 1992 reading of the College Entrance Examination Board's Advanced Placement (AP) Studio Art general portfolio submissions and discussions with experienced judges of portfolios that provoked particularly diverse ratings.

## Background

A coherent framework has evolved over the past century for inference from multiple-choice tests, in which "the tendency to get items like these right" constitutes an operational definition of competence. Classical test theory and item response theory (IRT) help us construct tests, interpret response patterns, and quantify the weight of evidence in terms of this proficiency. Values and judgments enter the process, to be sure, from the very decision to gather and interpret evidence about competence in this way, and continuing with the choice of topics to cover and the crafting of individual items. But all this takes place *before* students ever see the test. The objectivity of "objective tests" refers simply to the virtually uncontested agreement among observers as to whether, under prespecified testing conditions, a student marks predetermined correct options.

A distinguishing characteristic of performance assessment is that the student's response is no longer so simply and unambiguously classified as right or wrong; judgment is required *after* the response has been made. Judgment is a crucial link in the chain of reasoning from performance to inference about students. A task may stimulate creative or problem-solving thinking but to no avail, unless we can distill from a performance the critical evidence for the targeted inferences. It is thus essential to establish a common framework of meaning among judges—shared standards for recognizing what is important in performance and for mapping it into a summarizing structure. It is no less essential that the same framework of meaning be common to students and teachers as well. In addition to the issue of fairness (for students *should* know the criteria by which they will be evaluated), learning the standards for good work is part of what a course is meant to teach (Wolf *et al., op cit*).

A performance assessment system can be implemented in myriad ways. There are alternative options for scoring schemes, training activities, reporting strategies, specificity of scoring rubrics, offering of choices, and so on. We need to learn more about the consequences, costs, characteristics, advantages, type of evidence provided, and so on, of these alternatives, in order to build performance assessment systems that provide the right *kind* of evidence for our purposes, with the required *weight* and *coverage* of evidence, expending the right level of *resources*. This presentation proposes no definitive answer about which configuration is "best;" indeed, different configurations are almost certainly better suited for different purposes and different contexts. Our focus is rather on ways of gathering and interpreting information about critical aspects of the operation of a performance assessment system, in order to monitor key features of its operation, to identify local anomalies, and to highlight aspects that might suggest the need for change.

Our study employs two distinct perspectives, "statistical" and "naturalistic," that are used in tandem to analyze and improve large-scale assessment systems; that is, too large for all the interactions to be easily viewed and grasped by all the participants. Our example, the AP Studio Art General Portfolio Evaluation, involved 24 judges (i.e., "readers," in AP terminology)

evaluating nearly 4000 portfolios in 1992, with 13 ratings on each, for a total of over 50,000 judgments. This approach reflects contemporary thinking about quality control in industry (e.g., Deming, 1980) and accords with Ronald Fisher's (1973) theory of "acceptance" testing. A statistical framework is established for analyzing data, in order to quantify typical and expected sources of variation (in our case, students, readers, and portfolio sections). Variability is present in any system; within a statistical framework, typical ranges can be modeled. For a system that is "under statistical control," the major sources of variability have been identified and observations tend to follow regular patterns. Quantifying these patterns is useful, first because it explicates the uncertainty for final inferences (in this case, students' final scores) associated with aspects of the process, and allows effects on the process to be monitored when changes are introduced. Secondly, this framework allows one to identify observations that lie outside the usual ranges of variability, sometimes due to special circumstances that can be accommodated within the existing system, but sometimes that suggest changes to the system. This framework helps focus attention where it is most needed. For the statistical component of our project, we use J. Michael Linacre's (1989) FACETS program (described below).[1] Other examples with a similar perspective include Lunz, Wright, and Linacre (1990), Lunz and Stahl (1990), and Engelhard (1994).

Although statistical analyses tell us where to focus attention, they cannot tell us what to look for. These special cases are unusual precisely because the expected causes of variation do not explain them. Further insight at this point requires information beyond the factors the statistical framework embodies, illuminating the special reasons behind an anomalous rating or stimulating a new hypothesis about a previously unrecognized factor. Such investigations constitute the "naturalistic" aspect of our project, endorsing the crucial role of the insight of the people intimately involved in the process—in our case, the AP Studio Art readers, each of whom is, above all, a seasoned artist and art educator. We identified 18 portfolios that received discrepant ratings from two experienced readers in one of the rated sections. We discussed sections of each of these portfolios with two experienced readers to gain insights into the judging process in general, and into features of these particular portfolios that may have made rating difficult.

**Teacher Preparation and Instructional Design**

Advanced Placement Studio Art, established in 1972 and administered by ETS for the College Entrance Examination Board ("College Board" hereafter), is one of the nation's oldest large-scale portfolio assessment programs. The AP Studio Art Development Committee, composed of artists and art school, college, and high school art educators from across the nation, determines the portfolio requirements. Students submit portfolios containing samples of work they have completed in accordance with the requirements set forth by the AP Studio Art program. If a student's work is judged to exhibit accomplishments commensurate with those demonstrated in corresponding first-year college courses, the student becomes eligible to receive college-level credit for work completed in high school.

Each year the College Board publishes a document called the *Advanced Placement Course Description: Art*. In this document, the Development Committee delineates specific guidelines that students are to follow when submitting portfolios for evaluation. The Committee does not believe that a single AP studio art course can or should exist but rather encourages art educators to exercise their creativity in designing courses that will enable students to produce portfolios meeting the stated guidelines. As Askin (1985) notes, the AP Studio Art course "does not consist of a fixed body of ideas, but is affected by ongoing re-evaluations of both current and past art" (p. 7). Likewise, there is no one approved course outline and/or method of teaching. Teachers have a great deal of flexibility to create AP courses in studio art that will prepare their students for the portfolio assessment while simultaneously fulfilling district-level curriculum requirements. Another valuable publication, the *Teacher's Guide to Advanced Placement Courses in Studio Art* (Carnes, 1992), contains information about organizing and teaching an AP Studio Art course and includes examples of course outlines from existing art programs. The document is intended to provide teachers with some models they might consider adapting and/or expanding to meet the needs of their particular setting.

Throughout the history of the AP Studio Art program, the Development Committee has periodically revised the portfolio requirements. The portfolio's focus and emphasis has changed with the times (Askin, 1985). These changes are reflected in the course description the College Board publishes each year as well as in the full-color poster that students receive. The poster features exemplary art works from portfolios submitted to the AP Studio Art program in the previous year and provides a condensed version of the portfolio requirements. The three sections of the portfolio are defined, and guidelines are included for submitting works appropriate for each section.

During the year, the College Board sponsors a series of regional workshops to acquaint new teachers with the AP Studio Art program and with the portfolio assessment process and to expand the knowledge base of

experienced teachers. Typically, art teachers who have served as AP Studio Art readers and thus have direct experience with the program and with scoring issues conduct these workshops. During the workshops, the participating teachers have the opportunity to see sets of slides of student works. The assessment criteria are discussed, and the rating process is explained. By taking part in the workshops, teachers become more attuned to the assessment process and gain further insight into the portfolio requirements.

**Portfolio Requirements**

Students may elect to participate in two types of portfolio assessment in AP Studio Art: Drawing or General. We address General portfolios. The materials that the student presents for evaluation fall into three sections:

*Section A: Quality*

For this section, the student submits four actual works in one or more media. Students are asked to choose works that demonstrate a sense of excellence in art; works that "develop the students' intentions, both in concept and execution" (College Entrance Examination Board, 1994, p. 4).

*Section B: Concentration*

For this section, the student submits a series of up to 20 slides of related works or a film or videotape that show the student's exploration of "a personal, central interest as intensively as possible" (College Entrance Examination Board, 1994, p. 5). The concentration is made up of "a body of related works based on an individual's interest in a particular idea expressed visually. It focuses on a process of investigation, growth, and discovery.... Students are free to work with any idea in any medium. However, the concentration should grow out of a plan of action or investigation" (College Entrance Examination Board, 1994, pp. 5-6). The student prepares a brief written commentary to accompany the works, describing the nature of the concentration, the sources of ideas represented in the concentration, and the resources used. Figure 1 gives two examples of written commentaries.

*Section C: Breadth*

For this section, the student submits 20 slides of works to show a range of problems, ideas, media, and approaches he or she has worked with during the course. Section C consists of four subsections:[2]

> *Drawing*: The student submits slides of eight drawings, including works in which both line and tone are employed. The group of eight drawings is expected to show a range of expression and an exploratory use of materials.

## Figure 1

### Two Written Commentaries for Section B (Concentration)

Student 2938

**1. Briefly define the nature of your concentration project.**

The subject of my concentration is minimalist oriental landscapes particularly reminiscent of Chinese and Japanese landscapes. My fascination with landscapes and intense color use inspired me to emulate ancient oriental style along with minimalist simplification of form and clutter. I utilized their technique of depicting the serenity of nature through simple yet bold brush strokes and colors. My materials comprised of watercolors and airbrush. My series began with uncomplicated scenery and gradually building on to bolder use of form and color.

**2. What were the sources of the ideas represented in your project? What influences (things you have seen, heard, read, felt, or imagined) affected your work?**

My resources included numerous books such as *Art Through the Ages* and *Epochs of Chinese and Japanese Art*. I discovered from my research that the Chinese painted scenes from memory. Having learned that, I examined the scenery around my area during early morning and before sunset then attempted painting from memory. Particular artists that influenced me were 11th Century Chinese artist Fan K'uan and 15th Century Japanese artist Sesshu. I was extremely fascinated with the way Chinese and Japanese paintings were able to convey the beauty and essence of nature through simplified form and color.

**3. List any technical assistance you received or resources (if any) you used while working on the concentration.**

Critique sessions with my classmates aided me with my concentration. It helped me improve my forms and color usage. Extensive [sic] was also beneficial. It helped by making me realize the differences of the different periods of Chinese and Japanese art.

Student 3794

**1. Briefly define the nature of your concentration project.**

My concentration project grew out of a desire to explore *angularity* in a medium (clay/"wheel-work") which doesn't easily permit a graceful, lyrical expression of that term. I was initially intrigued by random geometric shapes depicted on rounded surfaces—often repeated on appendages of the main work—sometimes incised or emphasized by a glazing technique. Recently, I have begun to investigate those same geometric planes literally *piercing* one another as I have initiated an exploration of metal and wood. Reflective qualities and light(ing) have frequently been a concern as well.

**2. What were the sources of the ideas represented in your project? What influences (things you have seen, heard, read, felt, or imagined) affected your work?**

A preoccupation with natural science and physics was/is a source of the ideas represented in my (ongoing) project/work. "Machine-age" design principles have always appealed to me, as

## Figure 1 (continued)

has science fiction writing and non-prose addressing scientific concerns. Extreme manipulation of materials and mediums interests me. The application of *light* itself as a material interests me, as well.

**3. List any technical assistance you received or resources (if any) you used while working on the concentration.**

I received a wealth of technical instruction while attempting to produce sculptural lighting. Welding (basic), casting a resin/sand mold from a form, die grinding, tapping threads, and some electrical wiring were all taught to me by XX, Sculpture Professor at XX State University. I learned an extensive amount of information about chemical glaze composition as well as a variety of ceramic firing techniques from many local potters.... Owners of scrap-metal yards, as well, taught me much about the various properties of that medium.

*Color*: The student submits slides of four works in which color principles are the focal point.

*Design*: The student submits slides of four works in which the principles of visual organization are of central interest.

*Sculpture*: The student submits four slides that show two sculptures from two different views, illustrating both additive and subtractive techniques.

AP Studio Art portfolio assessment clearly differs from familiar achievement tests in many ways, not the least of which is the wide latitude of choice available to students as to subject, medium, and expression. The distinct requirements of the various sections and subsections are intended to ensure that all students will present evidence about these key aspects of artistic development, although the particular form that evidence takes may vary considerably from one student to the next.

**Standard Setting and the Rating Process**

The challenge readers face is to apply common standards to possibly quite different behaviors in different contexts. As we saw in Figure 1, one student's concentration focused on "angularity in ceramics," while another's dealt with an "application of techniques from traditional oriental landscapes to contemporary themes." It would be easier to compare students' performances if everyone were required to work with angularity in ceramics or oriental landscape, or a prespecified sample of topics—but these modes of evoking information provide no evidence about a crucial aspect of development as an artist, namely conceptualizing and realizing one's own challenges, in terms of scope, ambition, and match to one's own capabilities. How well the ceramics student might have fared with oriental landscapes is not directly relevant to the inference we are really interested in. What does matter, and what we must examine the fidelity of, is inference about the more abstractly defined qualities that should be evinced in any student's chosen concentration.

The "standard-setting" process for AP Studio Art is a two-stage process. During the first stage, the Chief Reader (the person who oversees the conduct of the portfolio assessment process) and Table Leaders (experienced readers who assist groups of four or five readers each) spend two full days examining, discussing, and selecting sample portfolios that demonstrate the full range of student ability the readers are likely to encounter. These selected portfolios exemplify the various points on the 0-4 rating scale that readers use to evaluate each portfolio section and subsection. The deliberations of the Chief Reader and Table Leaders, their discussions, and finally their decisions about which portfolios they will select as exemplars for training are analogous to the kinds of "standard-setting" activities often carried out in business and industry.

The second stage of the AP Studio Art standard-setting process involves assembling a team of 25 art school, college, and high school educators to evaluate the portfolios. Each year, about 15% of the readers are new to the AP Studio Art reading process, although all are familiar with the program. Activities included in this stage are designed to acquaint these readers with the reading process and to give them guided practice in applying the assessment standards. Readers discuss the sample portfolios in order to clarify meanings of each of the scoring criteria and to attempt to reach consensus in their usage of the various categories on the rating scale. The process of setting standards gives readers a sense of the range of portfolios they are likely to see, insights into special problems, and ideas about the ways various works correspond to the grading scale. The goal of standard setting is to have the readers internalize the standards, so that they can use the scoring rubrics in a consistent fashion. Additional information on standard setting and the rating process in AP Studio Art can be found in Askin (1985), Mitchell and Stempel (1991), and Mitchell (1992).

Before 1992, the meanings of rating scale values were conveyed only through training procedures, discussions, and feedback from Table Leaders and the Chief Reader (i.e., via an oral tradition). A draft written rubric was introduced and employed in the 1992 reading. This draft, shown as Table 1, was written at a level sufficiently general to be applied to all portfolio sections. In the 1992 reading, the readers evaluated its use. Following the reading, the AP Studio Art Development Committee discussed whether to maintain the general rubric or to introduce section-specific rubrics, and if so, the form each rubric should take. The committee revised the general rubric, and readers used the revised rubric for the 1993 reading. For the 1994 reading, section-specific rubrics were devised and piloted for the first time.

The readers rate the portfolios over a six-day period. Students' portfolios are identified by number only and are randomly assigned to readers. Readers work independently and do not discuss particular portfolios with each other. The readers evaluate the portfolios section by section, with section-specific training preceding the scoring of each section. Each reader is trained to evaluate all three sections of the portfolio. Section A is rated by three different readers. Sections B and C are each rated by two readers. Until 1993, a reader produced four ratings on each portfolio he or she rated for Section C, one each for the four subsections (i.e., drawing, color, design, and sculpture). Each portfolio thus received 13 ratings: three for Section A, two for Section B, and eight for Section C.[3] Ratings take the values of 1, 2, 3, or 4, but a score of 0 is occasionally assigned to a section if the student submits no work for that particular section or subsection.

The Chief Reader carefully monitors the ratings throughout the six-day reading. Whenever a portfolio receives ratings from two readers that differ by two or more points in a particular section or subsection, the Chief Reader reviews the portfolio to resolve the discrepancy. In a typical AP Studio Art reading,

## Table 1
## Draft Scoring Rubric (May 18, 1992)

The student is expected to meet portfolio requirements for each section. The following rubric describes work at each of the score points. It is not expected that the students' work will be equally strong in all the areas listed; within a score point, strength in one area may compensate to some degree for weakness in another.

4 Demonstrates *excellence* in understanding of
- use of art methods, techniques, and materials
- formal elements and principles of art
- visual language and art concepts
- imaginative, exploratory, and experimental approaches to visualization
- assessing quality and appropriateness of work for inclusion in portfolio

3 Demonstrates a *strong* understanding of
- use of art methods, techniques, and materials
- formal elements and principles of art
- visual language and art concepts
- imaginative, exploratory, and experimental approaches to visualization
- assessing quality and appropriateness of work for inclusion in portfolio

2 Demonstrates a *moderate* understanding of
- use of art methods, techniques, and materials
- formal elements and principles of art
- visual language and art concepts
- imaginative, exploratory, and experimental approaches to visualization
- assessing quality and appropriateness of work for inclusion in portfolio

1 Demonstrates *little or no* understanding of
- use of art methods, techniques, and materials
- formal elements and principles of art
- visual language and art concepts
- imaginative, exploratory, and experimental approaches to visualization
- assessing quality and appropriateness of work for inclusion in portfolio

0 Works [meeting portfolio requirement] not submitted for this section

about 1 - 2 % of the ratings are found to be discrepant and need to be resolved (e.g., about 500 of the 50,000 ratings made in the 1992 reading differed by two or more points). Rare as they are, discrepancies play a crucial role in the current system. Bringing discrepancies to the attention of the Chief Reader soon after they were given improves quality at two levels. The first is to provide another well-informed examination of the work: "resolving" the discrepancy serves the immediate need of bringing additional resources to bear on a portfolio likely to need it. The second is to build the Chief Reader's awareness of broader patterns that may merit attention: observing frequent discrepancies involving the same styles, media, or readers can prompt discussions with individual readers or the group as a whole on applying standards to the kinds of works that are appearing. The logistics of the current system preclude the real-time evaluation of ratings in light of a portfolio's ratings on other sections or a reader's ratings of other portfolios.

**Weighted Total Scores and Reported Scores**

The 13 ratings each portfolio received were summarized into a single, weighted, total score, such that the three sections contribute equally. The weights used in 1992 were 2.0 for each of the three Section A ratings, 3.0 for the two Section B ratings, 1.2 for the two Drawing ratings in Section C, and .6 for ratings in the Color, Design, and Sculpture subsections in Section C. This process maps the readers' 13 0-4 ratings into a single 0-72 weighted total score. The final scores reported to students and secondary institutions range from 1 to 5, based on cut points determined by the Chief Reader on the 0-72 weighted total score scale. The cut points used in 1992 are shown as Table 2 (the rightmost column in this table will be discussed in the section on the FACETS analysis).

Additional descriptive information on the 1992 ratings appears in Table 3 (numbers of ratings in each category after resolution) and Table 4 (average between-reader correlations within rating categories, post-resolution inter-rater reliabilities that take into account the number of ratings for each section, and correlations of ratings across sections before and after correction for attenuation using the stated inter-rater reliabilites). This information is drawn from Bleistein, Flesher, and Maneckshana's (1993) unpublished summary report to the College Board. Correlations between two readers' ratings for the same section average about .65; correlations between their ratings for different sections average about .45. (We note in passing that the correlations among section A, B, and weighted-total-C ratings fit a one-factor model well; the eigenvalues of the disattenuated correlation matrix are 1.975, .285, and .113.)

**Table 2**

**Cut-points for 1992 AP Studio Art--General Portfolios**

| Reported Score | Weighted Score Range | 1992 Frequency | 1992 Percentage | $\theta$ Range* (in logits) |
|---|---|---|---|---|
| 5 | 47 - 72 | 648 | 16.8 | > 1.87 |
| 4 | 40 - 46 | 790 | 20.5 | .96 - 1.87 |
| 3 | 31 - 39 | 1416 | 36.8 | ⁻.43 - .96 |
| 2 | 24 - 30 | 741 | 19.3 | ⁻1.77 - ⁻.43 |
| 1 | 0 - 23 | 250 | 6.5 | < ⁻1.77 |

* Based on expected weighted scores corresponding to midpoints of cut-point gaps. For example, 46.5 is the expected weighted score for a student with $\theta$ = 1.87 logits.

**Table 3**
**Counts of Ratings by Category**

| Portfolio Section | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| A: Quality | 0 | 785 | 2129 | 837 | 138 |
| B: Concentration | 3 | 927 | 1733 | 962 | 264 |
| C-1: Drawing | 7 | 1109 | 1853 | 745 | 175 |
| C-2: Color | 8 | 1016 | 1884 | 825 | 156 |
| C-3: Design | 16 | 939 | 1933 | 842 | 159 |
| C-4: 3D | 70 | 1308 | 1740 | 634 | 137 |

## Table 4
## Classical Test Theory Reliability and Correlation Coefficients

| | Portfolio Section or Subsection | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C (total) | C-1 | C-2 | C-3 | C-4 | Weighted Total |
| Correlation among readers | .59 | .64 | .58 | .58 | .50 | .49 | .52 | |
| Number of Readings | 3 | 2 | 2 | 2 | 2 | 2 | 2 | |
| Reliability* | .81 | .78 | .78 | .73 | .67 | .65 | .68 | .89 |
| Correlations** | | | | | | | | |
| A: Quality | 1.00 | .52 | .66 | .66 | .57 | .55 | .42 | .79 |
| B: Concentration | .41 | 1.00 | .60 | .55 | .56 | .50 | .45 | .82 |
| C: Breadth | .53 | .47 | 1.00 | | | | | .80 |
| C-1: Drawing | .51 | .41 | .90 | 1.00 | .93 | .83 | .64 | .73 |
| C-2: Color | .42 | .40 | .82 | .65 | 1.00 | .93 | .67 | .66 |
| C-3: Design | .40 | .36 | .79 | .57 | .61 | 1.00 | .75 | .62 |
| C-4: 3D | .31 | .32 | .70 | .45 | .45 | .50 | 1.00 | .54 |

* This is inter-rater reliability for portfolio sections A (Quality), B (Concentration), and C (total) and for subsections C-1 (Drawing), C-2 (Color), C-3 (Design), and C-4 (3D), as obtained via the Spearman-Brown formula from the correlation among readers and the number of readings. For composite section C (total) and Weighted Total, correlations among portfolio subsections are also taken into account. C-1 thus has a between-reader correlation of .58 and a reliability of .73—this accords with Spearman-Brown—while C (total) has a between-reader correlation of .58 and a reliability of .78—there are more ratings, but not as internally consistent.

** Observed correlations among weighted scores below diagonal; estimated "true-score" correlations above diagonal, as obtained by corrections for attenuation using reliability values from above. No adjustments are made for restrictions of range.

These discussions provide insights into the rating process and particular difficulties the readers, and therefore the system, encounter. We reasoned that portfolios that were hardest to rate would spark the most interesting discussions about the challenges of applying common standards to unique works. We selected eighteen portfolios to discuss with readers. We focused our discussions around Section A for nine of the portfolios and around Section B for the other nine portfolios. Most of the portfolios had received discrepant ratings in the section of interest (i.e., either Section A or Section B), although a few portfolios were selected before rating because they featured an uncommon style or medium. The concentration on oriental landscape is an example of this kind.

We held the discussions in the evenings of the days the portfolios were rated. Two experienced readers who had not originally rated the portfolio section in question[4] met with a leader (Myford, Sims-Gunzenhauser, or Wilkins) to discuss the portfolio. We gave the readers the information in Appendix A as preparation. The leaders used the outline in Appendix B to guide discussions if necessary. Issues that arose naturally in the discussion of the work, however, held precedence over the outline. The discussions, which lasted between 15 and 30 minutes, were audiotaped and later transcribed. The following discussion summarizes and illustrates emergent themes. We begin with some general comments on standard setting. We next address the processes and the challenges of rating Section A, most of which also apply to the other sections of the portfolio as well. The special considerations of Section B are then discussed.

**Readers' Views on Standard Setting**

Standard setting in AP Studio Art is structured around the examination of actual student portfolios. Prior to the standard setting, the Chief Reader and the Table Leaders select portfolios that represent a span of accomplishment in the section being addressed (i.e., Sections A, B, or C). Portfolios are selected that show student performance at each scoring level (1, 2, 3, and 4) for that section. Some of these portfolios will show work that is typical of students at each scoring level (i.e., "anchor" portfolios). Other portfolios will be selected because the student's performance, while indicative of work at a particular scoring level, is unusual in some way (e.g., the student is working in a medium that readers would rarely encounter, the approach the student has taken departs from the norm, etc.). Still others will be selected because they illustrate recurring challenges the readers will face (e.g., evaluating works that are uneven in quality). The sample portfolios are selected to help readers gain a sense of the breadth and scope of work they are likely to encounter in the reading.

The sample portfolios chosen for standard setting are set out for display. The readers break into small groups, each group working collaboratively with a Table Leader. The groups move from portfolio to portfolio, each member of the group rating each portfolio independently. After viewing a selected sample of portfolios, the group stops to compare their ratings and to discuss the qualities

of the work that led them to give their ratings. They discuss qualities that, for example, distinguish the portfolios receiving 2's, though quite different among themselves, from the portfolios receiving 1's and the portfolios receiving 3's from those receiving 4's. They gain a sense of the range of possible performances that could conceivably fall into each scoring category. The rating scale categories contained in the rubric are intended to reflect levels of artistic accomplishment as gauged by the spectrum of accomplishment seen at the end of the first year of college study in studio art. Building a view of the kinds of qualities to look for from this perspective, despite the variety of ways they might be expressed, is the goal.

When there is disagreement about the rating of a particular portfolio during standard setting, the ensuing discussions are often lively and animated, but readers remain respectful of one another's judgments.[5] The Table Leader monitors the discussion, drawing readers' attention to the assessment criteria and reminding them of the necessity of referencing those criteria when making their judgments. During standard setting, readers often confront their own biases, preferences, and parochial attitudes and begin to realize that they must "leave this baggage at the back door," concentrating instead on the prespecified criteria in the rubric to evaluate the portfolios. This process, of scoring a sample of portfolios and then reviewing the ratings given, is repeated several times during standard setting so that readers have the opportunity to become comfortable with the rating process. Over time, readers learn the meaning of each rating scale category as they encounter more and more examples of actual student work. With practice, readers' use of the various rating scale categories tends to become more settled.

A number of "veteran" readers note that participating in the standard-setting process is worthwhile even if one has served as an AP Studio Art reader for a number of years. They often find that they learn something new when they participate in a standard setting session. For the experienced reader, the recalibration process serves as a valuable reminder of "what counts" when examining work in the AP Studio Art setting.

The judgment process readers use in the AP Studio Art setting differs significantly from the process they use in judging art at exhibitions. In one reader's words,

> See, the portfolio [assessment] doesn't work unless you have the criteria set before, because otherwise you just become judges. And ... if it's an open show, he [the judge] picks the things he likes with only his criteria. But if the criteria are set for the portfolio ahead of time, we have to subjugate that—our criteria.... And the criteria for this is this: By the end of a college freshman year in a college entrance program, is this the work that would come out? (3056A, p. 3)

The key difference between the two evaluation processes is that in a juried exhibition, the judges base decisions on their own preferences using their own personal criteria. By contrast, in the AP Studio Art setting, readers must examine

student work against a set of prespecified criteria. The major challenge of standard setting is for readers to learn to use this agreed-upon set of criteria, subjugating their own personal criteria if need be. The readers must learn to view student work through a common lens, and much time is spent training readers to use that lens appropriately and consistently.

Although all readers are experienced artists and educators, all inevitably have their own unique backgrounds: the kinds of students with which they have worked, the media and styles with which they are more and less familiar, and what they personally respond to. Readers perceive a major function of the standard-setting sessions to be to give them a shared framework within which to exercise their personal experience and judgment:

> If you've got somebody ... who deals just with students [from a private school] as opposed to somebody like me who comes from a school that has a part-time policeman because of the situations that we've had, you're talking about people from two different environments looking at something. I think all that comes into play. One of the things I think standard setting does is make that real to everybody because it gives us all a chance to relate to what your situation is. (1900B, p. 2)

For many readers the experience of participating in an AP Studio Art reading serves to broaden their perspective, giving them a better sense of "the big picture" by moving them outside their own sphere. They have the opportunity to examine student portfolios completed in a variety of settings. They see the products of different environments in which learning takes place, and they gain a deeper understanding of how environment affects the art students produce.

AP Studio Art readers bring to the reading a wealth of experience in various visual art settings. They recognize that when functioning as a reader they must leave behind their own personal preferences for working in a certain style and/or medium. The reading is not a forum for advocacy. As one reader explains,

> We all have such rounded backgrounds of going to museums, of going to history classes, of being in classes with students with a number of different kinds of motives and seeing the motives of our own students over a period of time that I think most of us have overcome that sort of pettiness that you might find if you were simply to become an advocate. I don't see the reading process as one of trying to perpetuate a manifesto. I do that in my own work, but I don't do it when I come to operate professionally at a university, and I wasn't trained that way either. (3800B, p. 4)

Standard setting stresses the need for readers to confront their own parochial attitudes, personal preferences, and biases and to recognize that they are to be avoided when evaluating student work. In the example below, a reader discusses bias against cartooning and the experience of learning to overcome that bias so that the reader could view cartooning in terms of the criteria

contained in the rubric:

> That's one of the things that strikes me about it is that a lot of people can't give up what they think should be. Cartooning was my big thing [i.e., difficult for the reader to overcome bias against cartooning]. And I have given 3's and 4's to robot heroes and all that stuff because they fit this rubric in terms of concept, technical quality, and all that kind of thing.... If nothing else, if no kid ever benefited from what I'm doing in the process, I certainly have. It has made me grow and has made me a lot more open to what kids want to do and to let them explore things. (1900B, p. 8)

Another reader speaks of the AP Studio Art assessment experience as one of "moving beyond bias" to ask questions of the student's work posed by the rubric (e.g., Does the student manage the materials well? Does the student manage the subject matter well?) Management of materials or of subject matter are more objective criteria than whether or not the reader "likes" the student's work. By focusing on the criteria listed on the rubric, the reader's judgment is "brought to a more clinical level," providing "something more concrete, more substantial to talk about." The reader elaborates on this point, again within the context of judging cartooning:

> With the kind of open arena that we have for art today, a true artistic statement can be made in any number of ways, so I try to let myself be open and aware of what this person might be trying to do and if they are doing it well within that certain style, whether it's something I 'approve' of or not.... For instance, one of the styles that come up frequently in these ratings is a comic book, a cartoon style.... I say, 'Are they using the space well? Are they being inventive with how they're doing this, or are they merely mimicking somebody and not doing a very good job of it?' You almost go down a whole checklist of descriptors—line, space, color, texture. I kind of drill myself. I put myself under the spotlight. (2192B, p. 5)

One of the real strengths of participating in the AP Studio Art reading process, readers noted, is that the process requires them to develop an attitude of openness when examining student work:

> You come into this and you have to open yourself up to seeing lots of other approaches to art.... Last year we went round and round a couple of times about exactly what design was.... And several people said, 'That's not my idea of design.' And for others it was, 'But your idea is not the only idea, and you've got to be able to put aside your own biases' .... I think that's one reason this particular aspect of the program functions so well because people do do that. (1900B, pp. 7-8)

A number of readers mentioned that prior experience critiquing students' work in their own classrooms had served as valuable preparation for carrying out the task of reading AP Studio Art portfolios. In many art classes, a significant portion of instructional time is spent evaluating students' work, through self evaluation and peer evaluation as well as teacher evaluation. When critiquing student work, art teachers must remain open to a multiplicity of approaches to art making. Teachers are challenged to consider a variety of ways in which a

student might choose to solve a problem. They become adept at devising and recognizing the legitimacy of multiple solutions to a problem. Additionally, teachers gain valuable experience in overcoming their biases and prejudices. They must constantly put aside their personal preferences to evaluate student work using criteria to which students can relate.

**Evaluating Section A (Quality) of the Portfolio**

Students submit four actual works for evaluation for Section A of the portfolio. The reader assigns a single 1-4 rating reflecting the extent to which the works demonstrated a sense of excellence in art (i.e., quality). Each rating generally takes place in a minute or less. We asked readers to describe how they arrived at a judgment of quality. One reader described the process used:

> I tend to scan first ... the works as a group. My eye might be attracted to particular works, and I might focus on those for perhaps longer than maybe one or two other pieces in that group.... And then it comes to the more difficult part to describe.... I think going through the upper or lower [register] is definitely a help—deciding whether the group of work, the body of works together, represents a higher or a lower range. That decision is really the important decision to make at that point. (0269A, p. 3)

After the reader decides whether the work is in the upper register (i.e., in the 3 to 4 range) or in the lower register (i.e., in the 1 to 2 range), the reader then makes a finer discrimination (i.e., If the work is in the upper register, does it warrant a 3 or a 4? If the work is in the lower register, does it warrant a 1 or a 2?) This two-step decision process is introduced in standard setting, and readers are encouraged to use it during the reading.

The experienced AP Studio Art reader has made thousands of judgments of the quality of works of art. As is true with many forms of highly developed expertise, with experience automates certain aspects of this judgment process. In one expert reader's words,

> When I first came, I was more surprised at the speed at which the readers could do this. And yet, once your mind is going and you're seeing these things, it just seems to come automatically. And I think it's just ... all the times you've done it before. (0047A, p. 2)

Another reader speaks of the process as the "whew" experience:

> As important as making the rating is we do it so automatically. There's a holistic sort of 'whew' as we go through this. (0269A, p. 2)

Past experience plays a critical role in automating the judgment process, a reader notes:

> The more I think about it, it's the hundreds and thousands of these that I've seen in the past in my own class ... that helps clarify everything so much more. I hadn't thought about it really until we started this, but I'm sure they all shoot

> through my head—just instantly whoosh on through, you know. Then, again, that's based on every day for twenty-three years looking at some art work. (0047A, p. 2)

*Challenges Readers Face*

Readers face a number of challenges when rating AP Studio Art portfolios. In the next section of our paper, we include a discussion of some of those challenges. Our intent is not to denigrate the AP Studio Art reading process. On the contrary, this is a large-scale assessment program that for over twenty years has struggled to find effective tools and techniques for meeting these challenges. We believe that the program has learned some important lessons through that struggle, and that others can learn from their experiences. Our purpose in discussing these challenges is threefold. First, we describe the challenges in order to highlight how the reader training program tries to deal with these challenges. A significant portion of reader training is devoted to making readers aware of these challenges and to providing readers with effective strategies to meet the challenges when they arise. Second, we see this discussion of challenges serving as a diagnostic tool. If there are portfolios that readers view differently, we could look to the discussion to provide some clues that might help us try to explain why the readers' ratings differ. Third, we hope that the discussion will serve as a means of alerting those who are setting up new large-scale portfolio assessment programs to the kinds of challenges their raters are likely to encounter, to the importance of exploring those challenges with raters, and to the need for developing effective strategies to help readers meet those challenges in their particular assessment program.

In our discussions, readers identified a number of challenges they face when evaluating Section A of the AP Studio Art general portfolios:

(1) *Resisting operating in the "empathy mode."* When functioning in this mode, the reader thinks in terms of, "What could I do with the student if the student were in my class?" The temptation is to put on one's teacher's cap and to rate the student's work in terms of potential. During training, the readers are cautioned to constantly keep in mind that they are to evaluate the work as it is presented, and that they should not base their judgments on what they think the student is capable of accomplishing.

(2) *Avoiding the "bounce effect."* Difficulties can arise when moving from one portfolio to the next. One reader explains the "bounce effect" in these terms:

> You come off a great portfolio, and the next one is not as good. Sometimes if you're not careful, you'll grade it down lower than it should be. You've got to clear your head. (2532B, p. 8)

The "bounce effect" can also work in reverse. If a reader sees a string of portfolios receiving 1's, then when the reader sees a higher quality portfolio,

the tendency may be to give the portfolio a higher rating than is justified (i.e., than it might otherwise have received had it not been rated immediately after a string of lower quality portfolios). Readers are made aware of the "bounce effect" during reader training and are cautioned to take time to clear their minds between portfolios so that their judgments will not be influenced by their previous ratings. (In statistical terms, the "bounce effect" can be modeled as non-independent residual terms in a rating model, related to order of presentation. Such effects can be detected and quantified with multiple ratings of work obtained under different orders of presentation.)

(3) *Lacking knowledge of the context in which a portfolio was prepared.* The student provides written responses to several questions about the work submitted for Section B of the portfolio. These responses provide the readers with some context for decision making. However, when judging Sections A and C, the readers have no student background knowledge upon which to draw. Several readers mentioned their frustration over not being able to talk to certain students directly about the work they had submitted. They felt that such conversations would help them answer crucial questions that they had about the student's work, and would enable them to feel more confident about their judgments, particularly in those cases in which they were vacillating between ratings. However, one reader noted that having knowledge of context sometimes makes the judgment task more difficult, not easier. In the reader's words,

> When you know the child, it colors your grading ... because you know how hard Johnny worked to get here or that Bill only worked half as hard, and his looks just as good. [Evaluating students in my own classroom] is twice as hard. (0047A, p. 2-3)

In this reader's view, knowing the student might make it more difficult to evaluate the work objectively against a set of predefined criteria. Extraneous factors that may have little or no relationship to the criteria might serve to cloud the reader's judgments.

(4) *Using all four points on the rating scale.* During training, readers are encouraged to make use of the full range of scale points available to them and not to shy away from using the extreme points (i.e., 1 and 4). In the actual reading, the Chief Reader monitors the readers' ratings. If a reader is found to be giving ratings that frequently are more than one point different from other readers' ratings of the same student's work, then the Chief Reader will call the reader aside to talk about this. Readers voiced a concern that this monitoring process might cause some readers to take the path of least resistance, encouraging them to "play it safe." The "safe" reader would tend to use the extreme points on the scale infrequently while overusing the middle points (i.e., 2 and 3). "Safe" readers are less likely to be singled out as causing discrepancies than those who use all four scale

points. As one reader describes the dilemma:

> [We] stay with the 2's and 3's and we're not as adventuresome in using 1's and 4's as we might ought to be.... I think that we would use the extremities of the grade scale if we weren't so concerned about discrepancies. We'd use them more. We'd feel freer and not so paranoid. I don't know how valid that idea is, but artists are so diverse. There are so many different temperaments that we have and we reflect. (2205A, p. 7)

If many readers overuse the middle scale points, then the end result is that scores will tend to pile up around the center of the score distribution while comparatively fewer scores will be found at either end of the distribution. Readers wanted to know whether there were certain readers who used the middle scale points more often than other readers and if so, whether it was the pressure of being singled out as discrepant that caused them to adopt this strategy. (We looked at this question in more detail in our FACETS analysis of the rating data, which is discussed in the second half of this report.)

(5) *Lacking background and experience in the medium or style in which the student is working*. Readers acknowledge that they are sometimes faced with the difficult challenge of having to evaluate a portfolio that is in a medium or style with which they have had little personal experience. One reader describes the experience of evaluating such a portfolio and the need to maintain an open mind:

> This [portfolio] seems fairly alien to me in my own training background and also in the way I teach my own courses, but it doesn't mean that I value it any the less since I'm really trying to build on individual capabilities, so I think this is just another sense of competencies. The house of art is so large. There are a lot of places in it, so there's no reason why this isn't as valid as something I might project or try to develop in my students. (1636A, p. 5)

The readers readily acknowledge the need to remain open to a variety of approaches. The problem seems to be that when readers encounter such a portfolio, they may feel that they do not have sufficient information to respond to it in any depth. Readers are less likely to have solid reasons to back up their judgments because they lack experience with the style or medium. As one reader explains, readers have problems dealing with portfolios that "are different when we don't understand the difference" (3056A, p. 10). The natural tendency is to back off because the reader wants to be able to explain why he or she gave the portfolio a particular rating, but the reader may not be able to find the words to express it. In one reader's words,

> I've been teaching art for fifteen years. I've looked at thousands of pieces and I'm still struggling with the words. I've had a lot of training in art history, and I've taught art history, so I'm as well trained as I could be probably to do this, but I still can't put the exact words on it. (3056A, p. 11)

When a reader encounters a portfolio that in some way seems foreign to his or her own training and experience, the reader's visceral response to the work is not necessarily negative. A number of readers noted that they typically take a lot of time to examine such portfolios and try to give the student "the benefit of the doubt." In training, readers are reminded to use the "common lens" when they view work that seems foreign to their own background and training, concentrating on the criteria contained in the rubric and relying on those criteria as the basis for making their judgments. They are challenged to "try to be objective and intelligent about appreciating a language that might be beyond the scope of their experience" (0269A, p. 5).

(6) *Having considerable background and experience in the medium or style in which the student is working.* Several readers commented that the reader who has worked in the medium or style the student presents may view the portfolio differently than the reader who is not as experienced in that medium or style. With considerable background and experience may come an inherent sensitivity to work in that style or medium and, possibly, a tendency to be more critical and/or more cautious when dealing with student work in that style or medium. A reader reflects on the role his training plays in judging photographs:

> I'm just cautious when I deal with photographs because I think I have an inherent sensitivity to them. That was my undergraduate training, so I'm perhaps more critical of them. I like to just take time to look at them. (2884B, p. 1)

*Challenging Portfolios to Evaluate*

Readers described categories of challenging portfolios that they frequently encounter when scoring Section A of the general portfolio. In standard setting, these special challenges are discussed so that readers will be made aware that these types of portfolios are particularly difficult to read. Sample portfolios that exhibit these challenges are pulled and used as examples for readers to score and then to talk about so that new readers will have firsthand experience in learning to deal with the challenges. Even after training, readers may use differing approaches to evaluate particularly challenging portfolios. When portfolios are brought before the Chief Reader for adjudication because of discrepancies in the ratings, it is often because the portfolio exemplifies one of the challenges discussed below:

(1) *Portfolios in which the student exhibits good ideas but doesn't have the technical capabilities to see the ideas through.* Reader training stresses the need to consider both the quality of the ideas and the technical skills needed to execute them. However, during the actual AP reading, it is not uncommon for some readers to place more of a premium on good ideas over technique while others place more of a premium on technique over good ideas. In

such cases, discrepant ratings may result that will need to be adjudicated by the Chief Reader.

(2) *Portfolios that "depart from the norm"* (i.e., portfolios that are daring or eccentric in approach, departing significantly from the "tried and true"). Readers seem to have different opinions regarding the ability of the AP Studio Art assessment system to accommodate portfolios that depart from the norm. Some readers feel that the system exhibits sufficient flexibility to acknowledge and reward rugged individualism. In one reader's words,

> I think that psychologically, socially we're respectful of people who depart from the norm and are successful in that. I think that's embedded in this whole activity—that kind of individualism that is here. So when we see something that is different, I automatically become more attentive to that.... In that respect, I spend more introspective time trying to assimilate what they're saying to me. So is it treated in a different way? No. Is it looked at in a way that's more individualized? Yes, but I don't think that can be separate from the other components of the evaluative system. I think it's just human nature to expect that. (0126A, p. 4)

Other readers harbor some reservations about the system's ability to reward truly daring and innovative portfolios that attempt to break boundaries:

> Reader: I keep wondering if there aren't some kinds of eccentric things that we don't pick up on because we're so tuned to the standards [i.e., the rubric]. ... I mean, you're always faced with that as a teacher. You know, you give an assignment, and sometimes it's really irritating when somebody comes in with something that's really off the wall. You don't have the patience to look at it and think about whether they have really seen the problem in a whole different way.
>
> Interviewer: So there may be a bent to *not* look at the exploratory? You can't appreciate the risk taker, the really, truly, out of the mold?
>
> Reader: Right. The imaginative. (2205A, p. 7)

In this reader's view, the initial draft of the AP Studio Art written rubric placed a premium on foundational skills and undervalued the importance of imagination. Subsequent revised drafts of the general rubric included attention to imaginative, exploratory, and experimental approaches to visualization in an effort to extend the focus of the rubric (see Table 1). Many studio art foundation courses for freshmen in colleges and universities tend to downplay creativity and imagination and focus instead on skill acquisition, the reader explained. Consequently, since the AP Studio Art reading is designed primarily to identify students who appear to have a firm grasp of foundational skills (i.e., would "place out" of these foundation courses), the initial draft of the rubric focused heavily on skills and technique while deemphasizing (or, at the very least, downplaying, in this reader's opinion) the importance of imagination. If readers differ in

their views of the importance of imagination when evaluating students' work, then their ratings of that work may differ, he felt.

(3) *Portfolios that contain pieces that are of uneven quality showing a range of ability—some weak, some strong.* When reading such a portfolio, the reader's response will often be, "How can the same person who did this be able to do this?" Readers described different approaches to meeting this challenge, such as "grading up" to the strong pieces, "grading down" to the weak ones, and applying a compensatory model in which the good pieces compensate for the weaker pieces. Others choose to concentrate on the stronger ones, using the exceptional pieces "to inform them as to how to read the other pieces that might not be quite as exceptional" (2192A, p. 6). One reader poignantly describes this dilemma:

> I think when you have four pieces and one of them is truly bad, it's easy not to see that piece. The other three carry the fourth. I think it's harder when you've got two and two, and sometimes even harder where you have one that just really knocks your socks off, and then you have three that you think, 'I don't think this same person did this.' (0269A, p. 3)

(4) *Portfolios that contain works in more than one style.* Generally, when presenting work for Section A, students tend to show one style or, on occasion, four (the "I-can-do-it-all" portfolio). The portfolio that attempts to show more than one style is hard to read because it is difficult to form a unified conception of the accomplishment of a single student. The student may appear very accomplished when working in one style but may be significantly less accomplished when working in another, so the works appear to be of uneven quality.

(5) *Portfolios that use media that readers rarely encounter in AP Studio Art readings* (e.g., calligraphy, pottery, photography, batik, video). Portfolios submitted that incorporate media not often seen in an AP Studio Art reading may have a higher potential for receiving discrepant ratings. When readers have little experience judging work in a particular medium, it is sometimes more difficult to know what the standards of quality are for work done in that medium. Several readers noted that when they encounter such a portfolio, they find they must step back from the portfolio to ask themselves what the relevant criteria are for judging quality in that medium.

(6) *Portfolios that reflect traditions other than the Western European tradition.* Portfolios submitted by students of art schooled in traditions other than the Western European tradition present special challenges, some readers believe. For example, they argue, if students of art have been schooled in an Asian tradition that embraces formal elements unlike those of the Western European tradition, then their portfolios may be qualitatively different, reflecting a different aesthetic. Whenever possible, sample portfolios reflecting traditions other than the Western European tradition

are selected to be discussed during standard setting so that readers have the opportunity to zero in on important attributes of the works. However, even with training, some readers may encounter difficulties in scoring such portfolios. As one reader explained,

> I find that coming especially from the Eastern cultures, because they have such a different aesthetic than we have—and theirs is as legitimate—I'm always worried whether I'm being fair. Am I being fair to someone who has had an experience I've never had? ... And yet I'm the one that goes along and judges [portfolios submitted by] the students from that culture. I have a problem with that. (2938B, p. 8)

Some readers disagree with the notion that different traditions reflect different aesthetics. They feel that the rubric is sufficiently broad in scope such that the works of all students, regardless of their cultural background and training, can be appropriately evaluated. In their view, the criteria included in the rubric are not culture bound but, rather, are culture fair.

Those who believe that different traditions reflect different aesthetics would question whether readers who have had little or no experience with a particular tradition can meaningfully apply the rubric to evaluate works that embody the aesthetic of that tradition. They would argue that if readers are not well grounded in the aesthetic of the cultural tradition, then they would not have the expertise required to assess that portfolio. The rubric might be culture fair but, in this case, the way the readers applied the rubric would not be since the readers would be unable to view the work through the lens of that cultural tradition.

*Using the Scoring Rubric*

Prior to the 1992 AP Studio Art reading, the meaning of each rating scale point was defined implicitly through examples and discussions; no written rubric was in place. For discussion purposes, the AP Studio Art Development Committee provided the 1992 readers with a draft written rubric shown as Table 1, a general set of guidelines meant to be applied to all portfolio sections. The rubric was intended to serve as a statement of valued aspects of accomplishment rather than as a specific algorithm to be applied in evaluating any particular portfolio. In our interviews, we discussed this particular draft and its utility.

Since 1992, the members of the AP Studio Art Development Committee have engaged in an ongoing dialogue on issues surrounding rubric development, trying to reach consensus on the language to be used in specifying evaluation criteria and on rubric formatting. In 1992 and 1993, readers used different versions of a generalized written rubric. In the 1994 reading, the readers experimented with section-specific rubrics for the first time. We anticipate that over the next several years the AP Studio Art Development Committee will continue the dialogue on rubric development and perhaps suggest experimenting further with rubrics in various formats. The committee's desire

is to develop high-quality rubrics that can be shared with teachers and students, so that the evaluation criteria the AP readers use to make their judgments can be made clear to others. By sharing the rubrics, they hope that studio art teaching and assessment can be brought into even closer alignment than they currently are.

Unfortunately, there is a paucity of literature to draw upon that can guide efforts to design or revise rubrics. Clearly, at this stage in research on rubric development, we have little indication of what works, what doesn't and why. While a number of large-scale assessment programs have been in the business of rubric development for some time, few programs have documented the process they use to devise rubrics, let alone described any attempts to experiment with various formats for presenting their evaluation criteria. In many cases, a program will adopt a specific approach to rubric development and will continue that approach with little change for an extended period of time. As long as the approach seems to "work," there seems to be little incentive to experiment with alternative approaches that might work better. This has not been the case in AP Studio Art, however. Over the last several years we have witnessed a healthy debate among readers over many basic issues related to rubric design (e.g., Do we need written rubrics? If so, should we use a generic rubric that we can apply to all sections of the portfolio, or should we use section-specific rubrics? What criteria should be included in the rubric(s)? What format should be used for presenting the criteria? How many scale points should there be?) It is against this backdrop that we offer comments readers provided about the utility of the generalized rubric they used in the 1992 reading as well as some of our own observations about that rubric. Perhaps this section of the report may be of some use to the AP Studio Art Development Committee as it continues its deliberations on these critical rubric design issues.

Readers noted that the generalized written rubric was particularly useful for new readers and for recalibrating readers year to year. They felt that experienced readers tended to refer to the rubric less frequently during an actual reading than inexperienced readers since experienced readers have, by and large, internalized the criteria from their years of experience in using the criteria to make judgments. For experienced readers, the written rubric most often functions as a quick reference. However, experienced readers noted that there were certain instances in which they found the written rubric particularly helpful. For example, one reader felt that the rubric was especially useful for dealing with difficult-to-score portfolios because it helped him define yet again the important issues that need to be taken into consideration:

> I do think that using the rubric becomes much more important the more uneven or the more challenging the group of works appears to be. It helps to align my thinking in the appropriate directions, I think. (0269A, p. 4)

If the experienced reader is having a hard time deciding between two rating scale points (e.g., between giving a 2 or a 3), then the written rubric is also

helpful since the rubric identifies the different breaks between excellent, strong, and moderate. In one reader's words,

> If it comes down to it, I can go back to this [the written rubric] as a kind of final straw to draw, and say, 'Well, I can't make up my mind between a 2 and a 3.' Then I pull out that key phrase—strong, moderate. Which one is it going to be? Then it helps me. In that respect it does. (2192A, p. 3)

The written rubric also functions as an aid for "clearing one's mind." It helps to ensure that the reader is focused on the scoring criteria and not on his or her own prejudices and biases:

> Reader 1: I think you also have to be aware of your own personal likes and dislikes and prejudices.... You have to know what they are—and I think we all are aware of what we respond to and what we don't respond to—and when you see what you don't respond to, you have to make absolutely sure that you're looking at the work itself: the formal elements, the visual qualities, the concept that the student is trying to [get across] . . .
>
> Reader 2: That's why the rubric would be good to have because you can start to clear your mind and make sure you're back focused on what you're supposed to be. Because, you know, a lot of people if there's a lot of blood and gore and cartoon figures chopping off these people's heads and things, well, it's hard for a lot of people to get back to. . .
>
> Reader 1: To get beyond that. (0047A, pp. 4-5)

When asked whether the generalized rubric adequately captured all the relevant criteria for evaluating Section A of the portfolio, some readers indicated that the rubric needed to be amplified to be more useful. In the next section of this paper, we discuss individual readers' suggestions for possible additions to a Section A rubric. Some of these are no doubt controversial. Not all readers would readily agree that they should be added to the rubric. We are not recommending that the rubric be revised to include these criteria; we are not in a position to know. Rather, our purpose in writing this section is to stimulate discussion among the readers and among the members of the AP Studio Art Development Committee about the adequacy of the current rubric. Readers' suggestions for additional criteria are included below:

(1) *Ownership or signature*—"that identity to your work that is your own" (0857A, p. 8); "the special touch they put into it" (0857A, p. 9); "some part of himself that is in the work that you can sense when you look at it, making a personal statement" (0857A, p. 10); "portfolios that show a sense of individualism" (0126A, p. 4).

While several readers advocated including ownership or signature as a criterion, not all readers would agree with that position. Some readers

harbor reservations about including ownership in the rubric if it is defined in terms of style:

> Interviewer: Two of the readers I talked with last night mentioned the importance of ownership, of being able to see in the student's work that it has a signature, that it was really 'their' piece, that it was something more than just an intellectual exercise, but that they had really made it their own. They felt that this wasn't covered in this [rubric].
>
> Reader: Are we talking about ownership [in the sense that] they take responsibility for it, or are we talking about ownership in terms of style? Those are two really different questions today. Gosh, I'd hate to think that my high school student as a senior is locked into a style that he's going to stay with for any period of time. (1900B, p. 9)

(2) *Passion or spirit*—"a sense that the student really, really enjoys what he's doing; that spirit that is evident in the work itself that gives you a sense of what the student is about" (0857A, p. 8). Passion in not something that can easily be taught, as one reader explained:

> We teach the technical skills. We can teach anyone to be able to render and model and compose. That spirit, or that passion, is something that's much harder to get from the student. One of the things that we try to do is to get the students that we're working with to go beyond merely copying what it is that they're looking at— be it a chair, or a table, or a tape player, or a photograph— to go beyond that and put some of themselves into it so they have committed which is shown in the way that he/she has modeled some of this, or has composed, or has built these shapes and put them together. (0857A, p. 8)

(3) *Exploration or risk taking*—a willingness to venture beyond the formulaic, exercising one's freedom to experiment, to break boundaries. A reader emphasizes the importance of risk taking:

> One of the things we talked about today was risk taking. I tend to weigh in favor of a person who's willing to make a strong visual statement, even if they don't have quite the capability of carrying it off because their aim is so high. And so I feel that I have to give credit to that person for aiming at a difficult target. It's like degree of difficulty would be in diving—that you give the score based on the content of the piece, as well as it's execution. (3056A, p. 2)

(4) *Creativity or imagination*—displaying idiosyncratic ways of responding to ideas or things that are unique or individualistic. In the dialogue which follows, two readers discuss the importance of imagination:

> Reader 1: I really think there's only one thing that we've left out of this set of rubrics. It's kind of accommodated by conceptual clarity, but I don't think that it is inclusive enough, and that's something to do with the idea of imagination.
>
> Reader 2: Yeah.

Reader 1: Imagination is not a real easy thing to talk about.... Our minds have idiosyncratic ways of responding to ideas or things that other people say or even ideas that we think of that are really unique and individual. Some of that is the most interesting thing about art. It's kind of funny. That's the thing that we search for as artists is that kind of idiosyncratic kind of value system nobody else on this earth has. (2205A, p. 8)

(5) *Conceptual importance*—the rubric presently focuses on whether the ideas the student is trying to express in his or her works are clear conceptually (i.e., show conceptual clarity), but the rubric doesn't address whether the student's ideas are important conceptually. This issue arose as two readers discussed a painting of a very large, melodramatic face. The two readers agreed that the student presented his or her ideas very clearly in the painting. There was little question as to what it was that the student was trying to say. However, the work did not go beyond conceptual clarity to "capture the observer" (1636A, p. 3). It lacked strength—"the sense of ideation that goes beyond just the bare facade of the face itself ... the quality of the structure that's underneath the face" (1636A, p. 3-4). The reader who suggested that conceptual importance be added to the rubric acknowledged that this criterion is somewhat controversial. Not all readers would agree that it should be included in the rubric, he felt. When assigning a rating, some readers would place a great deal of emphasis on whether the ideas the student is working with are conceptually important while other readers would place little or no emphasis on this criterion. Furthermore, it might be hard to train readers to use this criterion reliably when assigning ratings since there is often little agreement among readers in their judgments of which works make important conceptual statements and which do not, the reader acknowledged.

*Dealing with Discrepancies in Ratings*

Differences in opinion are natural to the arts, and AP Studio Art readers concur that it would be stifling to expect perfect agreement among readers:

> One of the things that I absolutely love about art is the fact that you are allowed to have different tastes, and that you're not obliged to like anything that anybody else has liked.... [F]rom that standpoint, I don't feel terribly defensive if [another reader] sees things differently than I do. And if he gives something a 1, and I give it a 4, that's not necessarily grounds for embarrassment. (2884B, p. 4)

In the system that has evolved in AP Studio Art, the Chief Reader who is in charge of the rating procedure carefully monitors the ratings given throughout the reading. Aside from resolving numerical differences, the discrepancy procedure sparks discussions about how to apply standards:

> And there are checks and balances ... because if you get somebody who ... gives it a 1 or 2 and somebody else has given it a 3 or 4, then somebody comes along who mediates that. And if I'm a person who is grading too low and

you're grading too high, they come along and say, 'Let's do another standard setting and let's see how you feel, and let's look at the big picture.' (1900B, p. 8)

The challenge is to encourage individual professional judgment, but within a shared framework of meaning from which substantial departures can be detected. In the language of quality-assurance, this is the distinction between the "natural variance" of a system under statistical control and the "special causes" that warrant attention—a distinction with which readers are well-acquainted:

Reader 1: ... But the other thing that I kind of differ with a little bit in terms of this system is I think we do too much reconciling of the discrepancies. I almost think they should be left to stand to show that there is a strong diversity of opinion. Maybe averaged out, as far as the score goes, but I just almost think they ought to be left because I don't think it's necessarily a mistake on either person's part.

Interviewer: It represents an honest judgment.

Reader 1: Yeah. Unless somebody is sleepy or fatigued, and that happens for sure.... In some cases where there's a 1 and a 4 given, I think there are probably good reasons for both.

Reader 2: Maybe it's something that needs to be discussed. The discrepancy needs to be discussed but let it stand. I'd be really interested in having those discussions. (2205A, pp. 6-7)

Readers view discussions around discrepancies as healthy. When there is a shared sense of trust among readers (as is true of the AP Studio Art setting), then readers can talk out their differences. Through discussion they can determine whether their own personal preferences and/or biases may have entered into their judgments. They are helped to step back from the work and to become less self-involved. When trust is established, there is little loss of ego in doing that, readers note. The discussion process helps the reader to see the work through another reader's eyes and to bring the readers back to focus on the substance of the portfolio.

Paradoxically, discrepancies provide, at once, one of the richest sources of information for improving a system and one of the greatest sources of anxiety for the people who constitute the system. Quality assurance systems in industry succeed to the degree that feedback about processes is managed positively and utilized constructively. Statistical modeling provides an understanding of patterns of variation that can serve both to allay anxieties and to focus attention where it is most needed. We shall have more to say about this in the section on Statistical Analysis.

**Evaluating Section B (Concentration) of the Portfolio**

Each student submits a series of up to 20 slides of related works for evaluation in Section B of the portfolio. The works should show the student's exploration of "a personal, central interest as intensively as possible" (Askin, 1985, p. 26). The concentration should grow out of a plan of action revealing the evolution of an idea. Students submit written commentary to accompany their slides in which they discuss the nature of their concentration, the sources of their ideas, and the nature of any assistance they received or resources they used (recall Figure 1).

Section B is unique in two ways. First, it challenges students to move beyond what they've become comfortable with to engage in sustained work, which is hard for them because they are used to short-term, sequential projects in classes in other subjects. Second, only here do the students speak directly to the readers through their written commentaries. Readers, as art educators, stress the importance in their own classrooms of the give-and-take, the conversations, the working and re-working, and the attempts to put nonverbal ideas into words. The written commentaries for Section B serve this function in a limited way:

> The only thing better in the reading would be ... to be able to talk to the kids and just say, 'What do you think if we suggested something like this?'—the thing we get to do in our classroom that we're never going to get to do here. And sometimes you just wish you could have a phone right here, and grab it, and call them and say, 'Your stuff is just great!' or 'You did this?' or 'Did you try that and that's not what we're seeing?' (0047A, p. 3)

During reader training the readers are instructed to focus upon three questions as they look at a set of works:

1) Is it a concentration?

2) Is it a quality idea?

3) Do the individual works exhibit quality?

The reader assigns a single 1-4 rating reflecting the extent to which the works succeed as a concentration. We asked readers to describe how they arrive at a judgment for Section B. Some readers begin with the written commentary:

> When I pick up the envelope, as I'm taking the slide sheet out, I'm usually reading the first couple of lines of their definition of their concentration before I even look at it because I just want to get a sense of what the student is articulating even before I look at the work. I don't read the whole thing before I do that, but I do try to read the first couple of lines just to get a feel for it verbally. And then, what am I seeing? (1984B, pp. 3-4)

Other readers will look at the slides first and then read the commentary. Still others alternate between reading and looking:

> Sometimes I look at the slides first, and then I just sort of try to take it all in and maybe even guess, 'Now what is this about?' and then read. Sometimes I make myself read before I ever look. I don't know. I just go back and forth. (0528B, p. 3)

How much of the commentary is read seems to differ from concentration to concentration. For some concentrations readers gain enough information in the first several sentences to understand what the student is setting out to do; for other concentrations, readers must dig deeper into the written commentary in order to grasp what it is the student is trying to accomplish. If readers have some doubts about whether the set of works is indeed a concentration, they will typically read more of the written commentary to help inform their viewing of the slides:

> I look at the first several sentences, and then I look at the slide sheet, and then I look back, and then I look at the slide sheet again and decide whether it's a concentration. If I can't figure it out from that, then I'll go further [into the written commentary]. (0553B, p. 5)

Readers indicated that the student's written response to the first question, the nature of their concentration, is foremost in their decision making. While readers may on occasion look at students' responses to the other questions (especially if there is some doubt about whether the works constitute a concentration, if further clarification is needed or the concentration is particularly difficult to understand), it is the response to the first question—especially the first 3-4 sentences—with which they are most concerned:

> Most of the time I just look for the 'What is this?' and then I look at the work. I really deal with that. Is the work responding to the statement? Is it indeed a response to that or not? Then, within the context of, 'Yes, it is a response,' then 'Is it quality work? Has the student done what they said they were going to do?' (2532B, p. 4)

*The Written Commentary*

In our interviews we asked readers to describe what it is they look for in the written commentary. As readers reviewed and discussed examples of strong and weak commentaries, they identified a number of concerns which we will attempt to summarize in this section of the report.

When preparing their written commentaries, students should clearly express their purpose or intent so that the nature of the concentration will be evident to the reader. Above all, students should strive for clarity, coherence, and consistency in their writing, readers noted. It is important for the student to be focused, to be able to work through an idea visually from beginning to end. Similarly, focus must be evident in the student's writing, readers felt.

It is particularly critical that students say in the first several sentences of the commentary exactly what they are attempting to do in the concentration.

The response to the first question should make clearly evident the idea behind the work. One reader explains,

> When I pick it up, I want it [in] journalistic style, and I want to have that first sentence be like a thesis statement of intent and pretty much target what they're after. And then I like to see it kind of filter down through the concept into the more, more finite, more defined areas here. And I want to see a relationship. If they say they are looking at Franz Kline and Jackson Pollock and Helen Frankenthaler, I want to be able to look up at the slides and say, 'Yeah, I can really see evidence,' and not just that they picked some names out of an art history book after having done work and say, 'Well, this looks like mine.' I want to see evidence of their learning from it. (2532B, p. 4)

The written commentary can be very short and simple but entirely adequate, readers explained. The commentary must illuminate the images, but it should not be a stream of consciousness piece. The student's response to each question should be well thought out and carefully planned. Students should write their response to the first question early in the school year so that they will have a target to aim for as they prepare their concentration.

Many students find the task of writing the commentary to be quite challenging. They may not have been trained to put into words what they are trying to accomplish visually. A reader highlights the importance of this activity:

> They've got to be able to bring that up to some conscious level because that's what this is for. This concentration is to make evident an idea—both in terms of verbally in the essay, and in terms of the work—and they should be consistent. (2532B, p. 4)

Initiating a project from within, showing evidence of thought processes at work in a visual form, and reflecting on the process in writing may be especially difficult for high school students who have had little or no training in how to write about art. Several readers stressed the need for students to practice writing their responses to the questions before they complete the form. Readers suggested having students make photocopies of the form so that they could practice, edit, and revise one or more drafts, before submitting a final draft with the portfolio.

*Challenges Readers Face*

Many of the challenges readers face in evaluating Section A are also evident when readers evaluate Section B of the portfolio (i.e., operating in an "empathy" mode, avoiding the bounce effect, using all four points on the rating scale, etc.). Additionally, many of the types of portfolios that prove challenging to score for Section A are also challenging to score for Section B (i.e., portfolios in which the student exhibits good ideas but doesn't have the technical capabilities to see the ideas through, portfolios that "depart from the norm"). However, there are a number of challenges that arise that are somewhat unique to Section B portfolios, such as judging slides rather than original works. Readers

identified several challenges related to working from slides:

(1) *Size and quality of the slides.* The slides submitted for Section B are not projected but rather are viewed in a slide sheet. Slide quality can be problematic. Sometimes the original works may be of very high quality, but the photographs of the works are poor quality and do not do justice to the works (e.g., A reader explains, "We had a problem with one today in the gymnasium where there were photographs of three-dimensional pots. They were beautiful pots, but the photographs weren't!" (3794B, p. 2)). Slides are small, and readers don't get a sense of the surface qualities of the works. Subtleties are often missed.

(2) *Three-dimensional concentrations.* Concentrations made up of three-dimensional works can also be problematic. Readers may have difficulty examining three-dimensional objects in a two-dimensional slide format since they cannot view the objects from multiple angles.

(3) *Slide order.* The ordering of the slides on the sheet can also make a difference in the way the reader reads the work. Some slide sets that students submit seem to have a sense of order; others don't. Ideally, the student should arrange the slides so that the reader can see how the student has evolved through the concentration, readers noted. The slides are easier to follow if they are ordered in some logical way (i.e., from weakest to strongest, from early to late, etc.).

When evaluating Section A portfolios, readers have no background information about the student or the work to inform their decision making. By contrast, readers do have access to some useful background information when scoring Section B portfolios. While the written commentaries help provide context for understanding the student's intentions, readers experience yet another set of challenges as they work with the commentaries. During standard setting, the various challenges are identified and discussed at length in order to provide readers with effective strategies for dealing with the challenges when they arise. Among the challenges related to incorporating the written commentaries that were discussed are the following:

(1) *The student produces a concentration but cannot describe it in words.* Sometimes the written commentary "gets in the way" of the reader's gaining an understanding of the concentration. The reader can sense from looking at the images that the student understands what a concentration is, and it is evident that the student has the ability to assemble a strong one, but the student cannot explain what he or she had done. The writing is weak and unfocused. In one reader's words, "You can see clearly that the student is just not verbal" (0528B, p. 3). This could create potential scoring problems if one reader places a high premium on the writing while the other reader does not. A reader must struggle with some critical questions when evaluating such a portfolio: How much of a role should the written

commentary play in my decision? Should I ignore the written commentary and concentrate only on the high quality images? When the commentary is weak, how much freedom do I have to read between the lines? A reader aptly describes this dilemma:

> There are many times when the commentary has absolutely nothing to do with the development of what the student is doing. Sometimes I read it, but then I have to decide for myself if for some reason he could not articulate what he was really working with verbally, but he could do it visually and just set that aside and not let it be a hindrance to my understanding of what he's doing. (3800B, p. 6)

(2) *The student writes a clear statement of purpose but is unable to carry out his or her intentions in producing the concentration.* In some instances, readers find that the student produces a clear, simple statement of purpose; but when the reader looks at the student's work, it is evident that the student was not able to actually do what he or she set out to do. The student lacked the technical capabilities to execute the ideas.

(3) *The written commentary is lengthy and rambles.* Occasionally, students will include additional pages to amplify their written responses to the questions that appear on the back of the envelope in which they place their slide sheet. In some cases, readers will agree that the extra length is warranted; the student may need additional space to describe adequately what he or she has done. In other cases, the additional pages may detract from the portfolio. Sometimes excessive writing seems to be an attempt to compensate for poor quality work, readers note. When a student writes a lengthy response to a question, the student frequently diverges from the stated purpose of the question including much irrelevant detail. The student ends up talking "around" the concentration rather than "about" it. If readers must dig through lengthy responses to find the information they need, then students may be doing themselves a disservice.

(4) *The commentary seems to have been written after the works were completed.* Sometimes students do not formulate a clear, coherent plan for producing their concentration. Instead, they complete assignments in their art course and then later try to find a common thread which connects their works. In one reader's words,

> Sometimes you have concentrations that are the invention of the commentary. The commentary has been made up after the fact and consists of a pretty long stretch.... [The student thinks] I did all these things. Now what do they have in common? How do I stitch them together? (0553B, p. 5)

It is the reader's responsibility to decide whether the works show evidence of having grown out of a plan of action revealing the evolution of an idea or whether the student "just threw a lot of different works together that happen to look similar enough to pull it off" (0553B, p. 5). In many cases,

the decision is a very difficult one to make. However, in standard setting the readers are reminded that the primary concern in judging concentrations is the art work: Do the works represent a concentration? The quality of the written commentary is to be only a secondary concern.

(5) *The student presents a high quality set of works but the written commentary is irrelevant.* In some instances, what the student writes about is not what the works themselves suggest the concentration is about. The reader must decide whether to ignore the written commentary or factor the irrelevant commentary into the decision-making process. One reader describes the dilemma:

> There are lots of cases where the quality of the concentration isn't relevant to the written commentary and what they say their concentration is about. One of our test examples yesterday was a body of work that was supposed to be about oppression in Jamaica. Very few of the people that looked at the actual slide images could read that that was their problem, that that was what they were about. They were a compelling body of clearly defined images and everybody felt they held together well, carried enough similarities and a feeling of a concentration and yet enough differences to feel like the person was really stretching himself expressively. And I'm not for a moment denying the importance of that person having had that notion of what he or she was about, but that wasn't what captured us based on what we considered. (0553B, p. 6)

(6) *The works appear to be loosely strung together until one reads the commentary.* For some portfolios it is critical that the reader read the written commentary in order to see the continuity in the work. Without the commentary, the works appear to have no connection; but with the commentary, the connection becomes readily evident.

For example, one of the Section B portfolios we examined contained sculpture, painting, and collage. A reader discussing this portfolio with us emphasized the importance of the written commentary for understanding this portfolio. In his words, "If you read the statement, then there's some consistency. If you don't read the statement, then it looks like he's jumping around from one thing to another and that it is not a concentration" (3600B, p. 1).

Another reader relates a similar experience of viewing a portfolio of figure drawings with and without consulting the written commentary:

> I saw one other portfolio that had a number of figure drawings in it, and they all seemed to have different concerns at first, just on first distant glancing. They seemed like they were all over the place. So I read the paragraph right away, and it said that they had a concern for light and light quality and how it defines the figure. Sure enough, you look back in there and they all had something to do with light—every one of them. Some of them were flooded with light and some were dim. And that's what made them look so different from a distance. There might be a light page with just a few marks on it, and then there's another one that's very darkly lit and some shadowy figure is coming out of there. (2938B, p. 7)

(7) *Even after reading the commentary there is some question about whether the body of work is really a concentration.* Readers differ as to the degree of cohesiveness they demand when deciding whether a body of work constitutes a concentration. If there is too much variety in a body of works, the "legalist" tends not to see the work as a concentration. By contrast, readers who are somewhat more flexible may be willing to stretch to try to see the connections between what may appear at first glance to be disparate pieces. One reader described the "legalistic" mind set:

> I could see it if this caused a discrepancy. Probably somebody was very legalistic in a way and said, 'Okay this [the written commentary] says movement. This is too varied. There are too many different kinds of things going on here. It is not a cohesive body of work. Therefore, in my opinion, it is not a concentration.' (0528B, p. 6)

(8) *The student changes the concentration midstream, or the concentration evolves into something quite different from the student's original intentions.* Sometimes a student will begin with a plan of action. After working with the plan for some time, the student may find that it isn't working. The student may alter the plan dramatically; and as a result of that change, some genuinely first-rate work develops. The student shows evidence in the written commentary of having evolved in his or her thinking and of having learned from the experience of changing the initial plan. This situation tends to be problematic for the legalistic reader who may not regard the resulting set of works as a single concentration.

Occasionally, the readers will encounter a situation in which the student's concentration has evolved beyond his or her initial intentions, but the student is not aware of that evolution. The reader as art educator can see the change, but the change is not reflected in the written commentary. In the conversation below two readers discuss how they handle this situation when it arises:

> Reader 1: They start out with this thing in their mind, and they do all this stuff. You as an outsider who are still connected to this person because you're right there, and you're not the reader, you can see this other stuff happening, but the kid can't because they've started with this mind set. We've often talked about this, 'Oh, if we could just pull or rearrange it in this pattern or take out these three things, this would be better.' But you can't do that, and you can't second guess them. You really do have to leave it alone.
>
> Reader 2: You have to take what they've said at face value. You really do. Some students lose points because what they've said here [in the written commentary] and what happens here [in the slide sheet] are just worlds and worlds apart, whether that's because they don't know the vocabulary or they don't understand what they've done. (1900B, p. 4)

*Using the Scoring Rubric*

While the readers agreed that the generalized written rubric prepared for the 1992 reading was useful, individual readers had a number of suggestions for revising the rubric to make it more appropriate for evaluating Section B portfolios. Many felt that it would be worthwhile to develop several more pages that carefully flesh out the requirements for scoring Section B. Such a document would be particularly useful for training new readers but would also be an important resource for experienced readers to fall back on when they encounter difficult-to-rate portfolios. The following issues arose in our discussions with readers about expanding the rubric:

(1) *Number of images.* The rubric should address the issue of how many images is enough. The point should be made that the number of images needed should be determined by the complexity of the work. If there are only three works in the concentration but each was very time consuming to produce, then that would be an appropriate number of pieces to include in a concentration. On the other hand, if there are only three works in the concentration and each required very little time to produce, then three is probably not a sufficient number of images for a concentration.

(2) *Weak commentary or no commentary.* The rubric should discuss how to evaluate a portfolio submitted by a student who cannot adequately express in words what he or she did visually. Can a student get a 3 if the commentary is weak? If the student submits no written commentary but the concentration shows high quality work, can the student get a 4? Is the written commentary required in order to get a 3 or 4? In our interviews with readers, we noted some genuine differences of opinion on these issues. All seem to agree that the focus is primarily on judging the art work and not the writing per se, but readers seem to differ in the amount of emphasis they place upon the role of the written commentary in their decision making.

(3) *Commentary has little or nothing to do with the images.* The rubric needs to address the situation in which the student's discussion of the nature of the concentration is totally unrelated to the images. (As one reader expressed the dilemma, "Well, if it says that it's about human emotion and it's got vegetables in it, what do you do?" (1900B, p. 3)) If the individual works making up the concentration are of high quality in this instance, can the student still get a 3 or 4?

(4) *Not all works need to be finished and polished for Section B.* Section A requires all "best" works, but the rubric should note that works that are of somewhat lesser quality can be included in Section B provided that they fit into the concentration and help to show how the student's thinking progressed. The focus in Section B is on assembling works that show genuine growth and development (i.e., the work gets progressively better over time). In

the discussion below, two readers discuss a portfolio that showed evidence of significant learning over time.

> Reader 1: [A former Table Leader] showed an example where the kid's concentration flopped. It was resin casting.... You can see in the slides where ... things didn't quite work. But he went through the whole process and talked about what he had learned, and they gave him a good grade on it because of what he had learned—not because it was some slick piece of work that everybody just 'oo-ed' over. . . . When I heard that, I thought that made it [the assessment process] valid. . .
>
> Reader 2: You know, that was a change for me, too. Always before I met [the former Table Leader] it was really important that my kids produce finished pieces that looked great and were solutions to problems. She really turned my thinking around. She said, 'What are we here for? We're here for those kids to learn!' (1900B, p. 8)

(5) *Concentration changes midstream.* The rubric should discuss how readers are to deal with the situation in which the student starts out with one idea for the concentration but at some point changes the concentration and goes off in a different direction. Can a student get a 3 or 4 if this occurs? A reader discusses this dilemma:

> You pick up two sets of ten slides, both of them complete within themselves, but it's not one concentration. I have problems deciding should that be a 2 because it's not one concentration, because they clearly stated in their narrative that in mid-year they switched and then they went on to narrate their second concentration. Should that be marked down for 2, or should you count it as two 1's? Or could they each be 3's? I think it would help me if there was more clarification on that kind of scenario. (2884B, p. 5)

(6) *A concentration of weak works.* The rubric should note that a student must do more than simply succeed in producing a concentration in order to receive a score of 3 or 4. A student can have a concentration and still receive a 1 or 2 if the overall quality of the individual works is weak. In one reader's words,

> Reader 1: This came up in a discussion I had today. It's not so much what constitutes a concentration. It's more, for some people, I think, 'Okay, it is a concentration. Therefore, I automatically am going to put it in the upper register, even though the drawing might be really not very good.' I ran into that in a couple of my discussions with people over discrepancies today.
>
> Reader 2: Normally, we cover that in the standard setting where we say that that's one of the considerations, but it shouldn't automatically bump it up into an upper level, if the quality of the work is substandard. It still needs to be a quality presentation. (0528B, p. 5)

(7) *High quality work but not a concentration.* In some instances, students will assemble a number of works that are technically very proficient, but the works do not constitute a concentration. How is the reader to deal with this situation? Can such a student receive a 3, or would this student only be eligible for scores of 1 or 2?

*Comments on the Discussions*

AP Studio Art readers inscribe numbers for row after row of portfolios, hour after hour. Aides gather and compile these tens of thousands of numbers, calling discrepant ones to the attention of the Chief Reader. Program staff pore over more numbers with the Chief Reader—correlations, distributions, cut-score implications. It's easy to get the impression that the system is mainly about numbers. It isn't. It's really about figuring out what developing as an artist means, and how to map student artists' unique and highly individual accomplishments into a common framework of standards. Our discussions with readers reveal some of the difficult and complex questions that such an endeavor inevitably raises—each of which is faced thousands of times each year, and each of which must be addressed continually if the numbers that leave the system are to possess any credibility or validity. We are grateful to the readers for their openness in discussing the challenges they routinely face. Fortunately, developing methodologies to support and inform such efforts is now a focus in test theory and quality assurance. It is to this topic we turn.

# The Statistical Analysis

## Overview

Extensive reader discussions for every section of every portfolio would be fascinating, but obviously AP can't hold them. Two readers discussing each of the six rating areas (Section A, Section B, and four subsections of Section C) of 4000 portfolios for twenty minutes adds up to over eight work-years.[6] However, statistical analyses of the principal outcomes of the evaluation process—the ratings themselves—can serve to characterize patterns in the data, provide a basis for monitoring changes, and highlight those aspects of the data and the process that do warrant closer attention.

FACETS models the probabilities of ordered-category ratings in terms of parameters for students, readers, tasks, and other "facets" of the observation setting that may be relevant, such as reader-group effects, time of day, day of the week, and student and reader background variables. Student parameters capture students' tendencies to receive high or low ratings; reader parameters, their harshness (i.e., severity) or leniency; task parameters, whether they are generally easy or hard to get high ratings on. We would expect a lenient reader to give a high rating to a generally high-scoring student on an easy task. Such expectations are formulated more precisely, in probabilistic terms, through the model (shown on the next page)—a simplified template that quantifies the main effects of the facets against a background of "typical" variation. As we shall see, typical variation serves as a standard for identifying unusual ratings, portfolios, readers, and so on. After outlining the model briefly, we structure our discussion around questions we explored with the FACETS output:

- How accurately are students measured?
- Do the AP Studio Art readers differ in the harshness with which they rate portfolios?
- How do differences in readers' harshness affect student scores?
- Are there aspects of readers' background and training that seem to influence the ratings they give?
- Do the AP Studio Art readers use the rating criteria consistently?
- Is it harder for students to get higher ratings on some sections of the portfolio than others? Can we calibrate ratings from the three sections as a single variable, or do ratings on certain sections frequently fail to correspond to ratings on other sections?
- Do some portfolios exhibit unusual profiles in ratings across Sections A, B, and C?

**The Model**

FACETS fits main-effects models to "logits," or the logarithms of odds of a given rating compared to the next higher one.[7] In this study, the model takes the following particular form: the log-odds of the probability that a student with a "true" proficiency of $\theta$ will receive from Reader $j$ a rating in category $k$ [denoted $P_{h,j,k}(\theta)$] as opposed to Category $k$-1 [denoted $P_{h,j,k-1}(\theta)$], on a rating scale for a Section $h$ with $K$ categories is modeled as

$$\ell n\left[P_{h,j,k}(\theta)/P_{h,j,k-1}(\theta)\right] = \theta - \xi_j + \eta_h + \tau_{kh}, \quad (1)$$

where $\xi_j$ is the "harshness" parameter associated with Reader j, $\eta_h$ is an "easiness" parameter for Section $h$, and $t_{kh}$, for $k=1,\ldots,K$, is a parameter indicating the relative probability of a rating in Category $k$ as opposed to Category $k$-1 for the scale of Section $h$. It follows that the probability of a rating in category $k$ on Scale $h$ for a student with parameter $\theta$ from Reader $j$ is

$$P_{h,j,k}(\theta) = \frac{\exp\left[k\left(\theta - \xi_j + \eta_h\right) + \sum_{s=1}^{k} \tau_{sh}\right]}{1 + \sum_{t=1}^{K} \exp\left[t\left(\theta - \xi_j + \eta_h\right) + \sum_{s=1}^{t} \tau_{sh}\right]}. \quad (2)$$

(The numerator is understood to be 1 for Rating Category 0.)

Variation among *students*, as a main effect, is anticipated; some portfolios evidence higher levels of accomplishment than others. The estimated "measure" for each student, $\theta$, quantifies the tendency to receive high or low ratings, taking into account the particular sections, readers, and scales involved in that student's observed ratings. Variation among *readers*, as a main effect, indicates that some tend to be more harsh or lenient than others, across portfolio sections and students. The uncertainty about final scores this effect introduces can be reduced by improving feedback on the application of standards to readers individually or in training sessions, by adjusting students' scores to account for the effects of overly lenient or harsh readers, or by having certain portfolio sections rated by additional readers. Variation among *sections* would indicate that receiving high ratings might be easier or harder in some portfolio sections than in others. The designers of an assessment that is rated in multiple sections may have preconceived notions about which sections should be harder than others, and why. Finding unexpected parameter estimates here, one would ask whether they reflect unexpected yet valid differences in performance across sections, or if rating standards should be raised or lowered in certain sections.

In line with the "naturalistic" component of our project, FACETS also highlights particular reader/student/section ratings that are unusual in view

of the main effects—for example, a low rating from a lenient reader to a generally high-scoring student. Equation 2 gives modeled probabilities for the various rating categories, corresponding to estimates for each combination of parameters for particular students, sections, readers, and rating categories. Residuals, or differences between actual ratings and the expected value of these predictions, can be studied individually, or combined in terms of mean-square fit indices to examine patterns associated with particular portfolios, sections, or readers.

We don't expect perfect matches between observed ratings and modeled predictions. There will always be some variation in judgments, even of informed and experienced readers; hence the use of a probabilistic model such as FACETS, rather than a deterministic model in which the ideal is zero error variance. The objective is to flag for further investigation those parts of the data that seem to be out of synch with the usual patterns of variability. We used the FACETS "infit" mean-square index to gauge whether the residuals for a given student, section, or reader are typical (around the expected value of 1), unusually small (much less than 1), or unusually large (much larger than 1).[8] For example, a low fit index for a reader, signaling less variation than usual, can indicate that this reader tends to "play it safe," by using extreme points on the rating scale less often than the other readers. A high fit index for a portfolio can indicate that a student has performed unevenly across sections, or that readers disagree in their ratings of a particular section.

It is worth emphasizing that eliminating all variation and discrepancies from the system is not the goal. An ideal system would exhibit some degree of informed disagreement among experienced readers *within a common framework of meaning*. Our concern lies with disagreements caused by lack of understanding of the task or ambiguities about standards of evidence. Problems of this type that manifest themselves as anomalous ratings can often be rectified by refining the rating task—by defining criteria more sharply, clarifying values in discussions of problematic performances, or introducing separate scales to disentangle confounding aspects of performance. Clues for specific courses of action may be found in the principals of the anomalous ratings, leading us back to the works themselves and to the readers of hard-to-rate portfolios.

**Results**

We fit several variations of the model shown as Equation 1 to the rating data. Some included additional parameters for reader background variables. Some addressed ratings from only a single section. Others used a single set of rating category parameters for all sections. We also compared analyses before and after the discrepancies had been resolved by the Chief Reader. We focus on the results of the model shown as Equation 1 *before* resolution, and point out additional or contrasting findings from other analyses when relevant.

We begin with a quick overview of the estimates of main effects. Figure 2 depicts the estimates for students, readers, and portfolio sections along the

logit scale. Bands representing the final 1-5 AP reported scores are also shown; we calculated these in terms of the expected values of 0-72 weighted total scores for points on the $\theta$ scale (refer back to Table 2). The "sections," "readers," and "categories within sections" estimates were specified to center around zero. Note that the students are much more spread out than the readers, with standard deviations of 1.44 and .31 respectively. To give a feel for the logit scale, Table 5 maps logit-scale Student Measures to probabilities for Section A ratings. The most likely rating for a student whose measure is between about -1 logits and 1.5 logits, a range that contains about half of the students, is 2, which is in fact the most common rating.

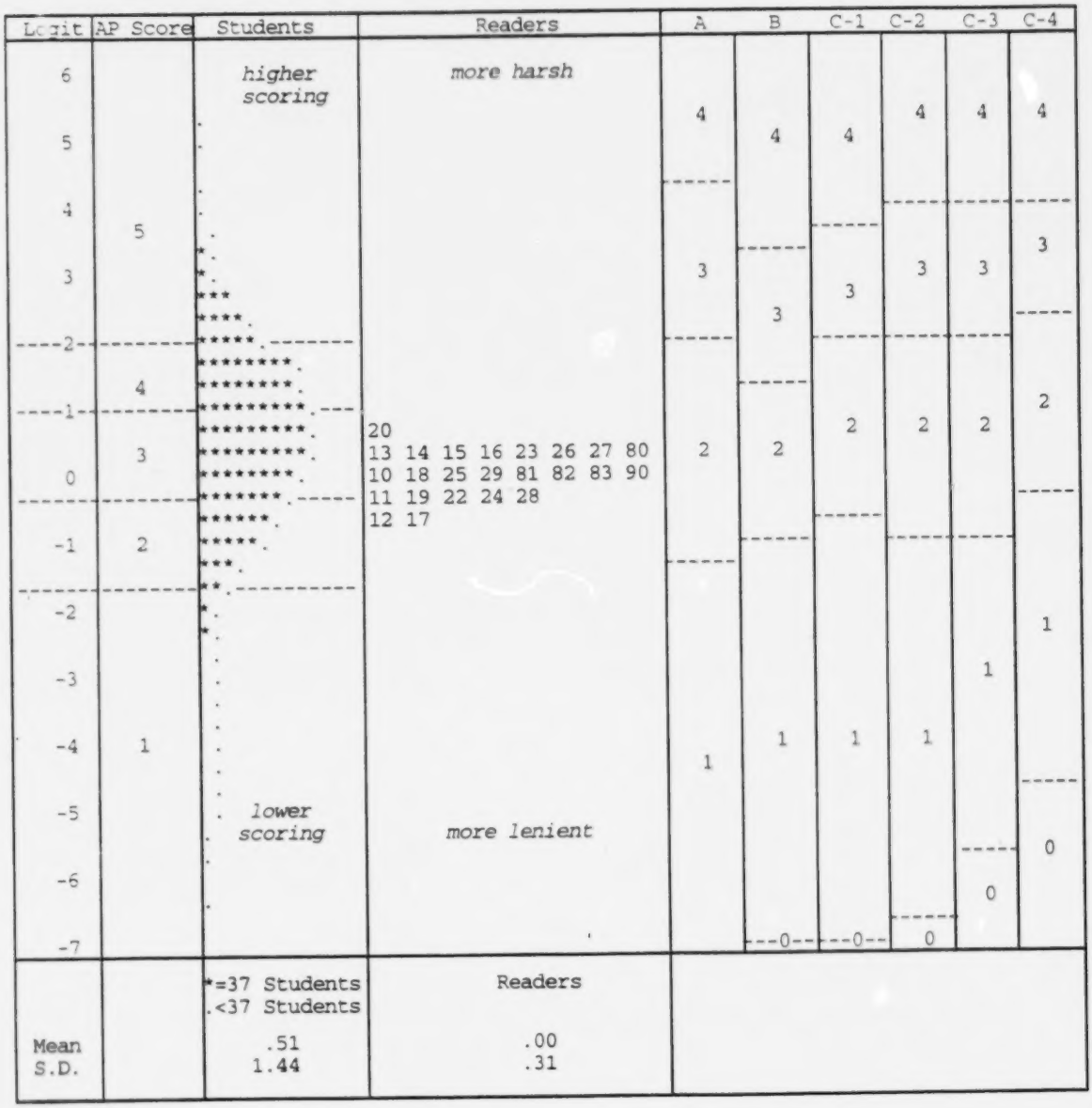
*How accurately are students measured?*

The standard errors for Student Measures average .47. Although we have only one estimated measure per student, we can generate two hypothetical measures per student with this amount of "noise" to illustrate aspects of the accuracy of students' AP reported scores. Figure 3 plots the simulated measures against one another, with dashed lines indicating the AP reported score boundaries.[9] The correlation between the two hypothetical Student Measures, a familiar measure of reliability, is .90—which is high by familiar educational testing standards, even for multiple-choice tests. The visual counterpart of this number is that the diagonal swarm of points is fairly compact. Because the cut points for AP reported scores are fairly close to one another in relation to the standard errors of measurement (averaging about 2⅓ standard errors apart), however, two statistically equivalent measures of the same student would fall in the same AP reported score category about 62% of the time—which seems low by familiar standards. The visual counterpart is the number of points inside the squares along the diagonal, which indicate identical AP reported scores for both observations. The bottom-line inference for each individual student is whether she made the cutoff for class credit in the institution she will attend. The proportions of consistent decisions are 95% for the AP reported score 2-or-above distinction (or 2+), 88% for 3+ (the most common decision rule among institutions), 86% for 4+, and 76% for AP reported score 5 alone.

*Do the AP Studio Art readers differ in the harshness with which they rate portfolios?*

This question addresses an important aspect of fairness: Students should not be disadvantaged if they happened to be rated by harsh readers nor unfairly advantaged if they happened to be rated by lenient readers. Examining discrepant ratings is not an appropriate method for dealing with this problem. Two harsh readers may agree in their ratings of a student, but without knowing that the two readers rate significantly more harshly than other readers, one would have no reason to question these ratings if the only quality control procedure in place involved identifying and resolving discrepancies. Finding that readers differ substantially in the degree of harshness exercised can suggest a need to address such differences in reader training, or to consider the feasibility of

**Figure 2**
**Estimates of Main Effects**



## Figure 2 (continued)
## Interpreting the "Map"

The "map" shown as Figure 2, based on a similar figure produced in the FACETS output, enables one to view all facets of the analysis in one figure, summarizing key information about each facet. It highlights results from more detailed sections of the FACETS output for students, raters, and other facets of the design.

The FACETS program calibrates the readers, students, and portfolio sections (i.e., the three facets of our analysis) so that all facets are positioned on the same scale. That scale is in log-odds, or "logit" units, which, under the model, constitute an equal-interval scale with respect to appropriately transformed probabilities of responding in particular categories. Having a single frame of reference for the aspects of the rating process facilitates comparisons within and between facets. The logit scale appears as the **first column** in the map.

The **second column** displays the five-point scale that the Advanced Placement Studio Art program used to report scores to students. The program converts the readers' 13 0-to-4 ratings for each student into a single 0-to-72 weighted total score, then determines cut-points to map these weighted totals into the 1-to-5 scale that AP Studio Art uses to report scores.

The **third column** displays the estimates of students' proficiency on the portfolio assessment—single-number summaries on the logit scale of each student's tendency to receive low or high ratings, across portfolio sections and readers. In FACETS terminology, these are "Student Measures." These Student Measures are ordered with more proficient students appearing at the top of the column and less proficient students appearing at the bottom of the column. Each star represents 37 students, and a dot represents less than 37 students. For any given dot or star, one can readily determine what a Student's Measure on the logit scale is by looking across the figure to the first column, and what the corresponding reported AP score is by looking across to the second column. The Student Measures range from -6.5 logits to 5.5 logits, about a twelve-logit spread. These measures appear as a fairly symmetrical mesokurtic distribution, looking something like a "bell-shaped" normal curve—although this result was in no way preordained by the model or the estimation procedure. Skewed and multi-modal distributions have appeared in other applications of the model.

The **horizontal bands of dashes that run across the "Logit," "Reported AP Score," and "Students" columns** denote the reported-score cutoffs used in the 1992 portfolio assessment. We determined through the FACETS model and parameter estimates those $\theta$ values which had the cut-points as their expected weighted-scores. For example, look first at the "Reported AP Score" column and the "Students" column. The students who are symbolized by the stars and dots below the lowest horizontal band of dashes have measures from -7 to -1.77 logits, and have an expected reported AP score of 1 on the assessment. The students who are symbolized by the stars and dots above that line but below the next horizontal band of dashes have measures in the -1.77 logits to -.43 logits range, and an expected reported AP score of 2. (See the last column of Table 2 for the range of logit values which correspond to each scale point on the AP Studio Art reported scale.)

The **fourth column** compares readers in terms of the level of harshness or leniency each exercised when rating students' portfolios. Because each section of each portfolio was rated by more than one reader, readers' tendencies to rate portfolios higher or lower on the average could be estimated. In FACETS terminology, these are "Reader Harshness Measures." In this column, each two-digit number identifies a particular reader, with more harsh readers appearing higher and more lenient readers lower. Reader ID numbers appearing on the same line signify readers who exhibited very similar levels of harshness. Reader 20 with a Harshness Measure of .5 logits was the most harsh while Readers 12 and 17 with Harshness Measures of -.66 logits and -.77 logits respectively were the most lenient. When we compare the distribution of Reader Harshness Measures to the distribution of Student Measures, we see that the distribution of Reader Harshness Measures is much narrower; Student Measures show an twelve-logit spread, while Reader Harshness Measures show only about a 1.29 logit spread.

The **fifth** through **tenth** columns—one for each rated section and subsection of the portfolio—show the most probable rating for a student at a given level on the logit scale (as expected from a reader with average harshness) in each of the portfolio sections and subsections. (See Table 10 for numerical values of these ranges.) The horizontal lines across a column, calculated from the section and category parameters, indicate the point at which the likelihood of getting the next higher rating begins to exceed the likelihood of getting the next lower rating. Looking *within* the Section A column, for example, we see that students with measures from -7 logits up through -1.24 logits are more likely to receive a 1 than any other rating; students with measures between -1.24 logits and

1.97 logits are most likely to receive a 2; and so on.[1] Looking *across* columns, we can determine the most likely rating on each of the six scales for a student of a given ability level. For example, a student whose measure was 3.5 logits was most likely to have received a rating of 4 on Section B and 3's on Section A and the four subsections of Section C (labeled in the map as C-1, C-2, C-3 and C-4).

The bottom rows of Figure 2 give the **mean** and **standard deviation** of the distribution of estimates for students and readers. When conducting a FACETS analysis involving judges, it is customary to "center" the judge facet (i.e., in AP terminology, the "readers" facet), or constrain the estimates to have a mean of zero. By centering facets, one establishes the origin of the scale. As Linacre and Wright (1994) caution, "in most analyses, if more than one facet is non-centered in an analysis, then the frame of reference is not sufficiently constrained and ambiguity results" (p. 27); this can be deduced from Equation 1. In our analyses, we centered the "readers," "sections," and "categories within sections" facets, but not the "students" facet.

[1] Readers used a rating of 0 to indicate that a student did not submit work for a particular portfolio section or subsection. No 0's were given in Section A (Quality), but 0's were given in Section B (Concentration) and all four subsubsections of Section C (Breadth).

**Table 5**

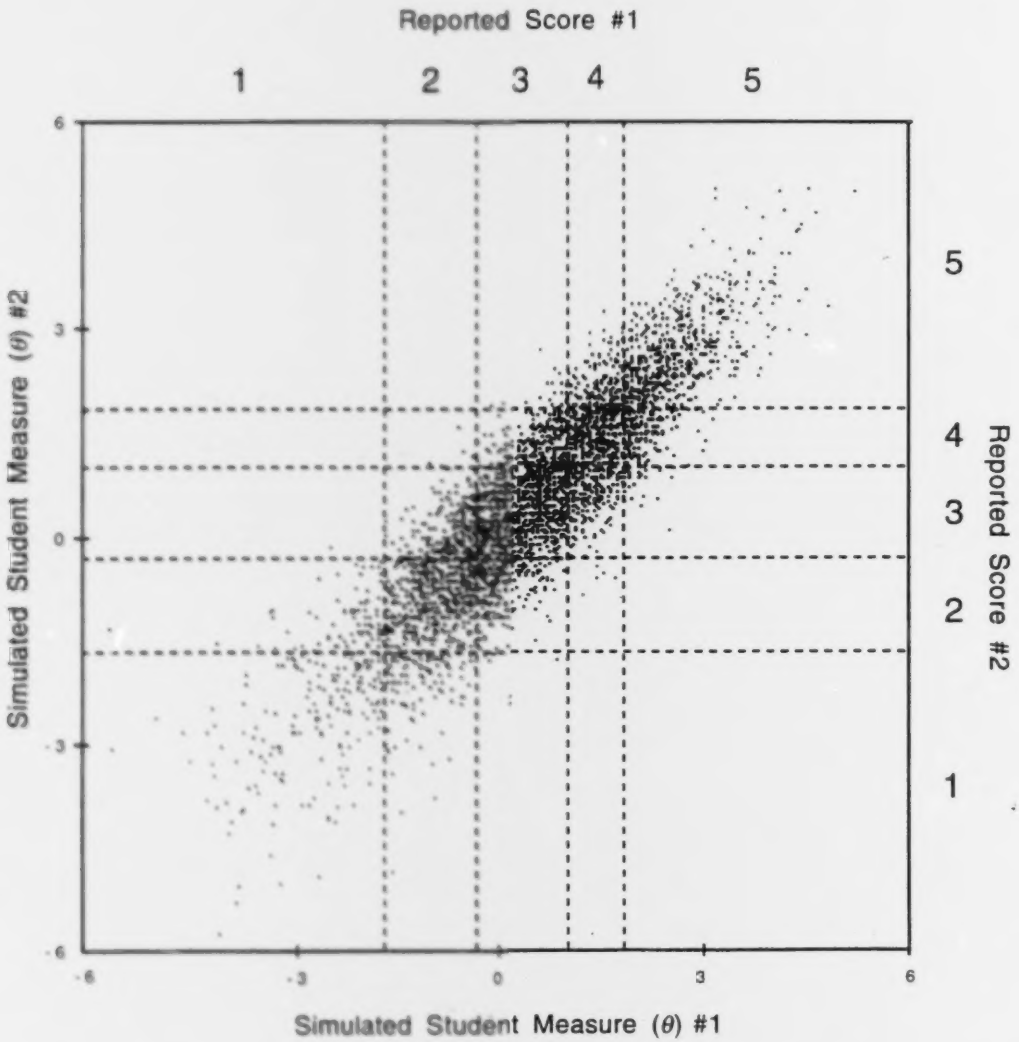
**Probabilities of Ratings for Students with Different Measures of Proficiency**

| Student Measures (in logits) | Probability of Rating Category | | | | Expectation | Most likely rating |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | |
| 6 | .00 | .00 | .11 | .88 | 3.88 | 4 |
| 5 | .00 | .01 | .26 | .73 | 3.72 | 4 |
| 4 | .00 | .06 | .46 | .48 | 3.42 | 4 |
| 3 | .00 | .20 | .57 | .22 | 3.01 | 3 |
| 2 | .02 | .45 | .46 | .07 | 2.58 | 3 |
| 1 | .07 | .66 | .25 | .01 | 2.21 | 2 |
| 0 | .20 | .70 | .10 | .00 | 1.90 | 2 |
| -1 | .43 | .54 | .03 | .00 | 1.60 | 2 |
| -2 | .68 | .32 | .01 | .00 | 1.33 | 1 |
| -3 | .85 | .15 | .00 | .00 | 1.15 | 1 |
| -4 | .94 | .06 | .00 | .00 | 1.06 | 1 |
| -5 | .98 | .02 | .00 | .00 | 1.02 | 1 |
| -6 | .99 | .01 | .00 | .00 | 1.01 | 1 |
| -7 | 1.00 | .00 | .00 | .00 | 1.00 | 1 |

Notes:

1. The Student Measures in column 1 are estimates of students' proficiency on the portfolio assessment--single number summaries on the logit scale of a student's tendency to receive low or high ratings, across portfolio sections and readers.
2. The mean and standard deviation of the Student Measures are .51 and 1.44.
3. Probabilities were calculated with respect to Section A effect and category parameters, and an average Reader Harshness Measure.
4. The "expectation" of a probability distribution is the average of the possible values, each weighted by the likelihood of its occurrence. For a student with measure of 6 logits, for example, the expectation is (1 x .00) + (2 x .00) + (3 x .11) + (4 x .88) = 3.88.

**Figure 3**
**Simulated Student Measures (θ) Illustrating Precision of Estimation for AP Reported Scores**



adjusting students' scores in accordance with the harshness or leniency of the readers who rated them.

FACETS produces a measure of the degree of harshness each reader exercised in rating portfolios, accounting for the various mixtures of high-rated and low-rated portfolios that the readers rated. These are estimates of the reader effect parameters $\xi_j$ in the logit units of Equation 1. Table 6 orders the AP Studio Art readers from most harsh at the top to most lenient at the bottom, in the column labeled "Reader Harshness Measure." To the right of each Reader Harshness Measure is the standard error of the estimate, indicating the precision with which it has been estimated. Other things being equal, the more observations an estimate is based on, the smaller its standard error. The Reader Harshness Measures range from -.77 to .52, a 1.29 logit spread.

A chi-square test of the hypothesis that the readers all exercise the same degree of harshness when rating portfolios has a value of 1540 on 23 degrees of freedom, meaning there is virtually no chance that the differences among estimated Reader Harshness Measures would have arisen from identically harsh readers. It is not surprising that this difference is statistically significant, because readers rated so many portfolios (about 2100, on the average); the differences are not so serious in their implications. As an illustration, Table 7 gives probabilities for ratings on an average section from an average student as rated by each of the readers, from the harshest to the most lenient. The most likely rating for all readers is a 2, with a probability around .7 in all cases. The harshest reader might say, "I'd probably give this portfolio section a 2, but if I really had to give it something else, it would more likely be a 1 than a 3;" the most lenient reader would agree that a rating of 2 was clearly most appropriate, but lean toward 3 rather than 1 if it had to be something else.

*How do differences in readers' harshness affect student scores?*

FACETS Student Measures take into account variation in reader harshness, adjusting for the leniency or harshness of the particular readers who rated each portfolio. In the logit scale, these are the $\theta$ parameters in Equation 1. They can be mapped back to the 0-4 observed-rating scale in several ways. FACETS calculates expected ratings for all students using a common standard reader harshness measure, calling them "fair averages". This is in contrast to the "observed averages" of ratings from the particular readers who happened to have rated them. In general, the rank ordering of students changes less frequently and less harshly as the readers vary less in their harshness and as more rate each student. In the same spirit, we calculated expectations of the 1992 Studio Art students' 0-72 weighted scores as averaged over all readers. Table 8 gives summary output for some representative students. The average difference between observed and expected scores is zero; the standard deviation is about .9. With an average distance of 8 points between cut points, the adjusted weighted scores of about 1 of every 20 students would move them up into the

## Table 6
## Reader Summary Table

| | Reader ID | Number of Ratings | Reader Harshness Measure (in logits) | Standard Error | Fit Mean Square |
|---|---|---|---|---|---|
| **Most harsh** | 20 | 2678 | .52 | .03 | .9 |
| | 15 | 3034 | .38 | .03 | 1.1 |
| | 14 | 2323 | .34 | .03 | 1.1 |
| | 16 | 2191 | .30 | .04 | 1.1 |
| | 80 | 2432 | .29 | .03 | .7 |
| | 26 | 2374 | .19 | .03 | .9 |
| | 13 | 2018 | .18 | .04 | 1.1 |
| | 27 | 1494 | .17 | .04 | 1.2 |
| | 23 | 1892 | .17 | .04 | 1.1 |
| | 83 | 1667 | .16 | .04 | .9 |
| | 18 | 1781 | .12 | .04 | 1.1 |
| | 90 | 546 | .09 | .07 | 1.1 |
| | 10 | 2493 | .00 | .03 | .8 |
| | 25 | 1849 | -.02 | .04 | 1.1 |
| | 81 | 2214 | -.03 | .04 | 1.1 |
| | 82 | 2862 | -.09 | .03 | .9 |
| | 29 | 2589 | -.11 | .03 | .9 |
| | 24 | 2278 | -.19 | .03 | 1.0 |
| | 28 | 2388 | -.21 | .03 | 1.0 |
| | 22 | 1209 | -.25 | .05 | 1.3 |
| | 11 | 2048 | -.26 | .04 | 1.0 |
| | 19 | 2518 | -.33 | .03 | 1.0 |
| | 12 | 1913 | -.66 | .04 | 1.1 |
| **Most lenient** | 17 | 1336 | -.77 | .04 | .9 |
| | Mean | 2088 | .00 | .04 | 1.0 |
| | S.D. | 552 | .31 | .01 | .1 |

## Table 7

## Probabilities of Readers' Ratings for an Average Section A Submission

| Reader ID | Reader Harshness Measure | Probability of Rating Category 1 | 2 | 3 | 4 | Expectation | Most likely rating |
|---|---|---|---|---|---|---|---|
| 20 | .52 | .20 | .70 | .10 | .00 | 1.90 | 2 |
| 15 | .38 | .18 | .71 | .11 | .00 | 1.94 | 2 |
| 14 | .34 | .17 | .71 | .12 | .00 | 1.95 | 2 |
| 16 | .30 | .17 | .71 | .12 | .00 | 1.96 | 2 |
| 80 | .29 | .16 | .71 | .12 | .00 | 1.96 | 2 |
| 26 | .19 | .15 | .71 | .14 | .00 | 1.99 | 2 |
| 13 | .18 | .15 | .71 | .14 | .00 | 2.00 | 2 |
| 27 | .17 | .15 | .71 | .14 | .00 | 2.00 | 2 |
| 23 | .17 | .15 | .71 | .14 | .00 | 2.00 | 2 |
| 83 | .16 | .14 | .71 | .14 | .00 | 2.00 | 2 |
| 18 | .12 | .14 | .71 | .15 | .00 | 2.02 | 2 |
| 90 | .09 | .13 | .71 | .15 | .00 | 2.02 | 2 |
| 10 | .00 | .12 | .71 | .16 | .01 | 2.05 | 2 |
| 25 | -.02 | .12 | .71 | .17 | .01 | 2.06 | 2 |
| 81 | -.03 | .12 | .71 | .17 | .01 | 2.06 | 2 |
| 82 | -.09 | .11 | .70 | .18 | .01 | 2.08 | 2 |
| 29 | -.11 | .11 | .70 | .18 | .01 | 2.09 | 2 |
| 24 | -.19 | .10 | .70 | .20 | .01 | 2.11 | 2 |
| 28 | -.21 | .10 | .70 | .20 | .01 | 2.12 | 2 |
| 22 | -.25 | .09 | .69 | .21 | .01 | 2.13 | 2 |
| 11 | -.26 | .09 | .69 | .21 | .01 | 2.13 | 2 |
| 19 | -.33 | .09 | .68 | .22 | .01 | 2.16 | 2 |
| 12 | -.66 | .06 | .64 | .29 | .02 | 2.27 | 2 |
| 17 | -.77 | .05 | .62 | .31 | .02 | 2.30 | 2 |

Notes:

1. Probabilities are calculated with respect to Section A effect and category parameters, and an average Student Measure.
2. The "expectation" of a probability distribution is the average of the possible values, each weighted by the likelihood of its occurrence. For Reader 20, for example, the expectation is $(1 \times .20) + (2 \times .70) + (3 \times .10) + (4 \times .00) = 1.90$.

next higher reported score category, and about 1 of every 20 would move to the next lower category; no students would move more than one category. The reader-effect adjustments are minor partly because the variance among readers is relatively small compared to variance among students, but more importantly because a random sample of seven readers contributes to the score of each portfolio. The overall levels of harshness of readers from student to student thus tend to average out.[10]

In the 1992 AP Studio Art data, the largest adjustment among the examples in Table 8 is for Student #3466—2.2 points upward on the 0-72 scale, from 30.0 to 32.2, which in fact crosses the reported-score category boundary from a high 2 to a low 3. The 13 Reader Harshness Measures that correspond to this student's ratings are as follows:x

.00, .18, .29, .17, .29, .19, .38, .19, .38, .19, .38, .19, .38.

The average is .25, compared to a mean of zero and a standard deviation of .11 over all students. Student #3466 had the unusually bad luck of being rated by readers consistently at or above average in harshness. While the impact on any one of the ratings is small, the effects in this case accumulate rather than cancel. This example serves as a reminder that even when the adjustments for individual students are negligible on the whole, small adjustments can be important for those students whose measures lie in critical cut-score regions. For institutions that use a reported score of 3 as a criterion, Student #3466's work would be eligible to receive college-credit with the adjustment but ineligible without it. When it is logistically feasible, flagging students near cut-scores for additional consideration, such as checking the average harshness of the readers who rated them, enhances the fairness of decisions.

The case for adjustment in AP Studio Art may also be less compelling because the readers use a 4-point rating scale. The most lenient reader in the AP Studio Art setting would tend to give proportionally more 3's and 4's, while the most harsh reader would give proportionally more 1's and 2's. In some other portfolio assessment settings, readers use scales having more points. If a student's portfolio were rated by very harsh readers who gave 1's and 2's or very lenient readers who gave 8's and 9's, then the amount of score adjustment needed would be greater than in AP Studio Art because the range of scale points employed would be much wider.

As we saw on the map of students, readers, and portfolio sections (Figure 2), the distribution of Reader Harshness Measures for AP Studio Art is much narrower than the distribution of Student Measures. However, it is not uncommon in a new performance assessment program for the range of Reader Harshness Measures to be as wide or even wider than the range of Student Measures. In AP Studio Art, a well established program, readers differ in terms of harshness, but the effects on individual students' scores while important are, nonetheless, relatively minimal (i.e., as we pointed out above, no students would

## Table 8

## Observed Ratings and Modeled Results for Selected Students

| Student ID | Section A | Section B | Section C | | | | FACETS results | | | Weighted Score (0-72) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | C-1 Draw | C-2 Color | C-3 Design | C-4 Sculpt | $\hat{\theta}$ | S.E. | Fit | Observed | Adjusted* | Change |
| 3751 | 1 1 1 | 4 4 | 1 1 | 1 1 | 1 1 | 1 2 | .49 | .45 | 4.5 | 36.6 | 36.3 | -.3 |
| 3347 | 2 1 1 | 1 1 | 3 3 | 1 2 | 4 3 | 2 3 | -.45 | .48 | 3.0 | 30.2 | 30.4 | .2 |
| 1377 | 3 4 2 | 1 2 | 1 2 | 1 1 | 1 1 | 1 1 | .30 | .47 | 2.3 | 34.2 | 35.1 | .9 |
| 2153 | 3 1 4 | 2 2 | 3 3 | 2 1 | 3 2 | 3 2 | 1.27 | .43 | 2.3 | 43.0 | 41.8 | -1.2 |
| 3366 | 1 1 1 | 2 2 | 3 1 | 2 2 | 3 2 | 4 2 | -.10 | .47 | 1.9 | 31.8 | 32.5 | .7 |
| 3360 | 3 3 3 | 1 2 | 3 3 | 1 1 | 3 2 | 3 1 | 1.09 | .44 | 1.8 | 40.8 | 40.5 | -.3 |
| 2712 | 3 3 4 | 2 2 | 1 2 | 1 2 | 1 2 | 1 1 | 1.16 | .44 | 1.7 | 40.4 | 41.0 | .6 |
| 106 | 2 2 2 | 1 1 | 1 1 | 1 1 | 1 1 | 4 2 | -1.07 | .51 | 1.6 | 26.4 | 26.9 | .5 |
| 1667 | 2 1 2 | 2 2 | 2 2 | 3 2 | 4 3 | 3 3 | .62 | .45 | 1.4 | 37.6 | 37.2 | -.4 |
| 35 | 2 1 1 | 2 3 | 2 2 | 1 2 | 1 1 | 2 2 | .05 | .47 | 1.4 | 33.2 | 33.5 | .3 |
| 2041 | 3 3 2 | 2 4 | 3 4 | 2 3 | 2 4 | 3 3 | 2.62 | .40 | 1.3 | 52.6 | 52.7 | .1 |
| 2538 | 3 3 2 | 1 2 | 2 2 | 2 3 | 2 2 | 1 1 | .74 | .46 | 1.3 | 36.4 | 38.0 | 1.6 |
| 2061 | 1 2 3 | 3 3 | 1 3 | 1 3 | 3 2 | 2 3 | 1.30 | .43 | 1.1 | 43.2 | 42.0 | -1.2 |
| 3063 | 2 2 3 | 2 1 | 2 2 | 1 2 | 2 2 | 1 2 | .12 | .47 | 1.1 | 33.8 | 33.9 | .1 |
| 852 | 2 1 3 | 2 2 | 2 2 | 3 2 | 2 3 | 3 3 | 1.01 | .44 | 1.0 | 38.4 | 39.9 | 1.5 |
| 682 | 1 2 2 | 3 2 | 1 2 | 2 1 | 1 2 | 2 2 | .11 | .46 | 1.0 | 34.6 | 33.9 | -.8 |
| 3785 | 1 3 3 | 3 3 | 2 2 | 3 3 | 3 2 | 3 3 | 2.00 | .41 | .9 | 47.0 | 47.5 | .5 |
| 1267 | 3 2 3 | 2 2 | 4 2 | 2 3 | 2 2 | 2 1 | 1.51 | .43 | .9 | 42.4 | 43.6 | 1.2 |
| 508 | 2 1 1 | 2 1 | 1 1 | 1 2 | 1 1 | 1 1 | -1.73 | .56 | .8 | 23.6 | 23.7 | .1 |
| 3466 | 2 2 2 | 1 2 | 1 1 | 2 1 | 3 2 | 2 1 | -.16 | .48 | .8 | 30.0 | 32.2 | 2.2 |
| 812 | 2 1 1 | 2 2 | 1 1 | 1 2 | 2 1 | 1 1 | -.85 | .50 | .7 | 27.2 | 28.1 | .9 |
| 2281 | 2 2 2 | 3 3 | 2 3 | 2 2 | 1 2 | 2 2 | 1.15 | .43 | .6 | 42.6 | 40.9 | -1.7 |
| 982 | 2 2 2 | 2 2 | 2 1 | 1 1 | 2 1 | 1 1 | -.24 | .48 | .5 | 31.8 | 31.7 | -.1 |
| 1143 | 2 2 2 | 3 2 | 3 2 | 1 2 | 2 2 | 2 2 | .85 | .44 | .5 | 39.6 | 38.8 | -.8 |
| 72 | 2 2 2 | 3 2 | 2 2 | 2 2 | 3 2 | 2 2 | 1.03 | .44 | .4 | 39.6 | 40.0 | .4 |
| 134 | 2 2 2 | 2 2 | 1 2 | 2 2 | 2 2 | 1 1 | .03 | .47 | .2 | 33.6 | 33.3 | -.3 |
| 1041 | 2 2 2 | 2 2 | 2 1 | 2 2 | 3 2 | 2 2 | .59 | .46 | .2 | 35.4 | 37.0 | 1.6 |
| 910 | 1 1 1 | 1 1 | 1 1 | 1 1 | 1 1 | 1 1 | -3.73 | .89 | .1 | 18.0 | 18.1 | .1 |

* Expected weighted score calculated with respect to each of the 24 readers, then averaged.

move more than one category). The variance between readers in this program is small, and each portfolio is rated by seven readers so individual reader harshness/leniency effects tend to cancel out to some extent. Adjusting students' scores for reader harshness does not substantially improve the overall reliability of the ratings in this program (see footnote 11).

By contrast, when a new assessment program is instituted, the biasing effects of reader harshness may be quite substantial if the variance between readers is large; ratings may depend more on who gives them than on the relative quality of the work! In this situation, scores adjusted for reader harshness are unarguably fairer to students than their unadjusted scores. Consider the Pittsburgh portfolio assessment program (LeMahieu, Gitomer, & Eresh, in press) and the Vermont program (Koretz, Stecher, Klein, & McCaffrey, 1994). These are both large-scale portfolio assessment programs in which only two readers rate each portfolio. The case for score adjustment is stronger in situations like these if the personnel in charge of these programs wanted to make high-stakes inferences about individual students, since the chances of drawing a preponderance of readers from the same end of the harshness continuum are much greater when only a few readers rate each portfolio. Having all readers rate all portfolios would eliminate the problem, since the average reader effects would be the same for all students. However, the cost associated with instituting such a complete judging plan is often prohibitive for large-scale assessment. In general, the fewer the readers who rate a portfolio, the greater the likelihood that there will be a subsequent need for score adjustment if there are sizable differences in harshness among the readers in the reader pool. If the variance among readers is large compared to the variance among students, then calibrating the readers and adjusting students' scores for reader harshness may substantially improve the reliability of the ratings, since the noise associated with this systematic source of variation in the ratings will have been removed (Braun, 1988; Houston, Raymond & Svec, 1991).

*Are there aspects of readers' background and training that seem to influence the ratings they give?*

We asked each AP Studio Art reader to complete a questionnaire that solicited information about his/her training and experience (see Appendix A). When we designed the questionnaire, we worked with the program's administrators to help us identify the "telling" background factors, those that were most likely to influence how a reader approaches the rating task. For example, administrators were interested in whether readers who taught at the college level produced ratings that were significantly different from readers who taught at the high school level. That is, were the college instructors any more harsh in their ratings than the high school teachers? We also compared the ratings of those who taught a number of art courses, such as drawing, painting, sculpture, and photography, to those who taught only one or two such courses. We were interested in finding out whether breadth of teaching experience made

a difference in how readers approached the rating task. Did those who taught a number of arts courses, and presumably were familiar with the criteria for judging art work in a variety of media, rate any more harshly or more leniently than those who specialized and were perhaps not as familiar with the criteria for judging work in more than one medium?

The administrators in charge of this program viewed this kind of analysis as critical from a quality control perspective. If the analysis of the rating data indicated that readers differed widely in the degree of harshness they exercised when rating students' portfolios, then the administrators would like to pinpoint those aspects of readers' backgrounds that may have been responsible for such differences. An analysis of background factors could yield practical information that those in charge of reader selection and training could use when working with readers.

We coded the questionnaire responses and used the information as background variables in another FACETS analysis, replacing the single reader parameter in Equation 1 with main effects for eight reader-background variables. Only two had a statistically significant impact on reader harshness, and the magnitudes of these effects were negligible. We found that (1) readers who had experience teaching at the high school level rated slightly more leniently than those who had no high school teaching experience, and (2) those who had 1-3 years of experience as AP Studio Art readers rated slightly more leniently than those who had served for more than 3 years. By and large, reader harshness could not be predicted from the background variables we studied.

*Do the AP Studio Art readers use the rating criteria consistently?*

Finding that some readers used the rating criteria inconsistently would be important quality control information to share with those who oversee the AP Studio Art program. Administrators in charge of the program would like to be able to identify those readers who are unable to internalize the rating standards and use them consistently when rating students' portfolios. Again, however, identifying inconsistent readers is only the first step. Improving the process also requires understanding the nature of their inconsistencies. In what kinds of situations does the reader exhibit inconsistent rating behavior? Administrators who oversee the assessment program need this detailed, diagnostic information to work with inconsistent readers in ways that will meet their individual needs.

FACETS produces measures of within-reader consistency for each reader, in terms of mean-square fit statistics—weighted and unweighted, standardized and unstandardized. As noted above, we concentrate on the unstandardized, information-weighted index, or "infit" as it is labeled in the printout. The expectation is 1; the range is 0 to infinity.

An infit mean-square value less than 1 indicates less variation than average in the reader's ratings—"too much consistency." Often the problem is

that the reader fails to use all the points on the scale. In the AP Studio Art setting, we were interested in finding out whether there were "safe" readers in the group who would tend to overuse the middle points on the 4-point rating scale (i.e., 2's and 3's) and avoid using the extreme points (i.e., 1's and 4's), so that they would be less likely to be singled out by the Chief Reader for having given a discrepant rating. A "safe" reader would show up in the analysis as having an infit mean-square statistic of less than 1.

A reader who is unable to distinguish between different aspects of performance and tends to give ratings that are very similar across rating scales designed to measure conceptually distinct aspects would also have an infit mean-square less than 1. For this reader, the rating scales do not function as separate, independent indicators of aspects of performance. This is less likely to be a problem in the AP Studio Art setting than in other settings, since the readers use a single scale to rate evidence from each section and subsection of the portfolio separately. The trouble is more likely to occur when a reader is asked to rate a student on a number of different traits from the same body of evidence at the same time. A low infit mean-square in these cases indicates that the ratings profile of the reader tends to be flatter than average.

An infit mean-square greater than 1 indicates greater than average variation in a reader's ratings, even after the particular portfolios involved in those ratings have been taken into account. This reader is not using the rating scales in as consistent a manner as the others. Perhaps there is a difficulty in developing a solid understanding of what a 2 is, or what a 3 is, and so on. In comparison with the patterns of the other readers, the inconsistent reader does not use the rating scales in the same way for all students and appears unable to maintain his or her personal level of harshness when rating students. (It is important to keep in mind the comparative interpretation of fit indices. High values simply indicate when readers are not using the scales like most other readers, not necessarily that they are using them incorrectly. If an experienced art educator and twenty statisticians rated the Studio Art portfolios, it would not be surprising to find the art educator had the highest fit value indicating extreme misfit with the statisticians' ratings. Her ratings could be at once least typical and most valid.)

There are no hard and fast rules for setting upper and lower control limits (Deming, 1975) for flagging reader infit mean-square values. Guidelines for a specific program will depend on both the nature of the program and the level of available resources. New programs and new readers will tend to show more extreme values than established programs and experienced readers; in any case, the most extreme values are more likely to signal special causes of variation, and therefore, are more likely to provide clues for improving the system. For some programs in our experience, upper and lower control limits of 2 and .5 were appropriate for initiating further investigation. For others with a more conservative approach, wishing to reduce further the variability among readers,

more stringent mean-square limits of 1.5 or 1.7 were used for the upper control limit and .7 for the lower control limit.

The reader infit mean-square statistics shown in Table 6 range from .7 to 1.3, which suggests that all 24 readers were internally consistent in their ratings. There is no indication that any of the them could be characterized as "safe" readers who overused the inner scale categories of the rating scales, nor that any used the rating scales inconsistently when rating students' work. Experience and training appear to have effectively prepared these readers to carry out their assessment task consistently.

Had any of the reader's infit mean-square values fallen outside upper or lower control limits, we would have turned first to the listing of individual misfitting ratings in the FACETS output for clues about the nature and sources of the inconsistency. The table of individual misfitting ratings inventories the most surprising or unexpected ratings, based on differences between observed values and modeled expectancies. It pinpoints the particular ratings that were unexpectedly high or unexpectedly low, taking into account the reader's overall level of harshness and the other ratings the portfolio received. If the contents of the portfolios are available, it can be determined if, say, a reader is overly harsh or lenient when rating artistic styles in which she has specialized knowledge or strong personal reactions. Table 9 shows part of the individual misfitting ratings table for our run, sorted by readers. We see, for example, that most of the discrepancies involving Reader 12 involve the 3-dimensional subsection of Section C. Do these portfolios employ a technique with which Reader 12 has special expertise and recognizes accomplishments (or lack thereof) that other readers miss? Or, conversely, are they works in a style or medium with which Reader 12 has less experience than most of the other readers? This detailed diagnostic information offers insights into the types of instances in which a reader exhibits inconsistent rating behavior, so that discussions or additional training could be targeted to that reader's special circumstances.

*Is it harder for students to get high ratings on some sections of the portfolio than others? Can we calibrate ratings from the three sections as a single variable, or do ratings on certain sections frequently fail to correspond to ratings on other sections?*

As for students and readers, the FACETS program produces measures of difficulty and indices of fit for each section of the portfolio, as well as parameters for category-within-section probabilities. Table 10 gives section fit indices, and lists ranges along the $\theta$ scale for each category within which that rating is most likely. These ranges are also depicted in Figure 2, the "variable map." Looking at the row for Section A, for example, we see that students with measures from $-\infty$ up through -1.24 logits are more likely to receive a 1 than any other rating; students with measures between -1.24 logits and 1.97 logits are most likely to receive a 2; and so on. (There is no range for a rating of 0 for Section A because no 0's were observed in the data for Section A.)

**Table 9**

**Individual Unexpected Ratings for First Four Readers**

| Reader ID | Portfolio Section or Subsection | Student ID | Observed Rating | Expected Rating | Residual |
|---|---|---|---|---|---|
| 10 | C-4: 3D | 1644 | 4 | 1.5 | 2.5 |
| 10 | C-4: 3D | 577 | 3 | 1.1 | 1.9 |
| 10 | C-4: 3D | 649 | 3 | 1.3 | 1.7 |
| 10 | C-3: Design | 2650 | 4 | 1.9 | 2.1 |
| 10 | C-4: 3D | 2650 | 4 | 1.7 | 2.3 |
| 10 | C-4: 3D | 2930 | 3 | 1.3 | 1.7 |
| 10 | C-4: 3D | 3015 | 3 | 1.2 | 1.8 |
| 11 | C-2: Color | 1351 | 1 | 3.2 | -2.2 |
| 11 | A | 2164 | 1 | 3.1 | -2.1 |
| 11 | C-2: Color | 2165 | 0 | 1.8 | -1.8 |
| 11 | C-2: Color | 2771 | 3 | 1.3 | 1.7 |
| 11 | C-4: 3D | 2854 | 3 | 1.3 | 1.7 |
| 11 | C-3: Design | 2942 | 0 | 1.9 | -1.9 |
| 11 | C-2: Color | 3055 | 4 | 1.7 | 2.3 |
| 12 | A | 1104 | 4 | 2.2 | 1.8 |
| 12 | C-4: 3D | 60 | 4 | 1.3 | 2.7 |
| 12 | A | 471 | 4 | 2.1 | 1.9 |
| 12 | C-4: 3D | 827 | 3 | 1.2 | 1.8 |
| 12 | C-4: 3D | 106 | 4 | 1.5 | 2.5 |
| 12 | C-4: 3D | 289 | 4 | 1.3 | 2.7 |
| 12 | C-4: 3D | 575 | 4 | 1.8 | 2.2 |
| 12 | C-4: 3D | 891 | 4 | 1.9 | 2.1 |
| 12 | C-4: 3D | 1207 | 3 | 1.1 | 1.9 |
| 12 | C-3: Design | 1349 | 2 | 3.6 | -1.6 |
| 12 | C-4: 3D | 2063 | 4 | 2.0 | 2.0 |
| 12 | C-1: Drawing | 2416 | 0 | 1.0 | -1.0 |
| 12 | C-4: 3D | 2709 | 4 | 1.5 | 2.5 |
| 12 | C-4: 3D | 2715 | 0 | 1.9 | -1.9 |
| 12 | C-4: 3D | 2828 | 3 | 1.2 | 1.8 |
| 12 | C-4: 3D | 3292 | 4 | 2.1 | 1.9 |
| 12 | C-4: 3D | 3348 | 4 | 1.6 | 2.4 |
| 13 | C-2: Color | 931 | 3 | 1.3 | 1.7 |
| 13 | C-4: 3D | 274 | 3 | 1.3 | 1.7 |
| 13 | A | 1974 | 4 | 2.0 | 2.0 |
| 13 | C-1: Drawing | 2568 | 3 | 1.3 | 1.7 |
| 13 | C-4: 3D | 2921 | 3 | 1.3 | 1.7 |
| 13 | C-4: 3D | 3164 | 3 | 1.3 | 1.7 |
| 13 | A | 3826 | 4 | 2.0 | 2.0 |

These ranges can be used to see whether it was harder for a student to get high ratings in some sections than in others. For example, the $\theta$ range for "2 is most likely" extends lowest for Section A, so it was easier to receive a 2 in Section A than any of the other sections. Section A also had the widest range within which a rating of 2 was most likely. But getting a 3 in Section A was just about as difficult as in any of the other sections (except for Section B, Concentration, which was somewhat easier than the other sections to get a 3 in), and Section A was the most difficult to get a 4 in. Such patterns mean little, in and of themselves; Section Difficulty Measures merely reflect empirical patterns that must be evaluated in light of intentions and expectations. When performance tasks are designed to vary from easy to hard, for example, a task meant to be easy but exhibiting a high difficulty measure demands scrutiny; experience does not match expectations. One would look for hidden assumptions, unclear directions, misapplied rating criteria, or unintended sources of trouble for students.

We noted when presenting descriptive statistics that the sections of the assessment were consistent with the definition of a single variable from a factor-analytic point of view, with Section C-4 (3D) somewhat less closely related to the rest. The section fit indices tell the same story; they are all within even tight control limits of .7 to 1.5. The highest value is for Section C-4 (3D); when inter-section profiles are found to be unexpected under the hypothesis of a single variable, the three-dimensional work in Section C-4 is more likely than the other sections to be the "odd one out." A single summary measure captures the essence of score profiles in most cases, however, with the understanding that the typical degree of variation around the central tendency is the norm.

What if we had found results that argued against constructing a single variable; say, a strongly multidimensional correlation matrix or high values for FACETS fit indices? This would suggest that a single score often fails to tell the whole story; that systematic kinds of profile differences are appearing among students who have the same overall summary measure. Further investigation may reveal either unintended differences, due to problems with students' understandings or readers' ratings, or intended differences, if ratings do indeed reveal valid consistent patterns of profile differences. In other AP programs, an example of the latter might be reflected in consistent differences between students' successes with multiple-choice and essay questions. Modeling could then proceed separately for the two sections for quality control monitoring and measuring students, with final scores as an externally imposed combination of two distinguishable aspects of competence—a combination based and defended on the grounds of values rather than measurement theory.

*Do some portfolios exhibit unusual profiles in ratings across Sections A, B, and C?*

AP Studio Art readers rate portfolios one section at a time, rarely seeing entire portfolios at once. When readers rate Section A (Quality), all the Section A submissions are arrayed in a large room for viewing, and readers spend a

## Table 10
## Portfolio Section Summary Table

| Rated Section or Subsection | Range of Student Measures in which this rating is MOST likely* 0 | 1 | 2 | 3 | 4 | Section Fit Mean Square |
|---|---|---|---|---|---|---|
| A: Quality | | -∞ to -1.24 | -1.24 to 1.97 | 1.97 to 3.95 | 3.95 to +∞ | .9 |
| B: Concentration | -∞ to -7.72 | -7.72 to -0.83 | -0.83 to 1.53 | 1.53 to 3.26 | 3.26 to +∞ | .8 |
| C-1: Drawing | -∞ to -6.78 | -6.78 to -0.58 | -0.58 to 2.03 | 2.03 to 3.61 | 3.61 to +∞ | 1.1 |
| C-2: Color | -∞ to -6.51 | -6.51 to -0.71 | -0.71 to 2.03 | 2.03 to 3.75 | 3.75 to +∞ | 1.2 |
| C-3: Design | -∞ to -5.77 | -5.77 to -0.87 | -0.87 to 1.89 | 1.89 to 3.79 | 3.79 to +∞ | 1.3 |
| C-4: 3D | -∞ to -4.25 | -4.25 to -0.17 | -0.17 to 2.26 | 2.26 to 3.80 | 3.80 to +∞ | 1.5 |

* These are ranges along the logit scale for each section within which a student would be more likely to receive that rating than any other rating. The upper end of a rating-category range and the lower end of the next higher rating-category range are equal; they are the sum of the FACETS section parameter estimate for the section and the category-within-section parameter estimate for the upper category. No range for a rating of 0 is shown for Section A because no 0's were present in the data for this section.

day or two rating all of them, and them only, using the scoring criteria for Section A. When all the Section A submissions have been scored, they are put away and Section B (Concentration) submissions are displayed for their ratings. Some readers were concerned that this section-by-section rating procedure might result in too much variability in ratings across some portfolios (i.e., a given student might receive high ratings on one section but low ratings on another). The current procedure for resolving discrepancies catches atypical variability in ratings *within* a given section of each portfolio, but there is no check on the degree of variability in ratings *across* the various sections of the portfolios. In some cases, variability in a portfolio's profile of ratings reflects genuine unevenness in student performance across sections. Although atypical, such profiles can be validated and defended. In other cases, the variability may stem instead from differences in the degree of harshness individual readers exercise when rating the student, or from difficulties in evaluating an unusual or unfamiliar work. These instances of inter-section discrepancy could be flagged for the Chief Reader's attention, as intra-section discrepancies are now.

If many portfolios exhibit large inter-section variability, the following experiment could be run: A large sample of portfolios would be rated twice, once in the current manner and once with readers still providing separate ratings for each section, but with the complete portfolio in front of them for supporting information about the student's skills, intentions, and styles. Finding little or no reduction in the number of portfolios with highly variable *ratings* across sections would support the hypothesis of truly variable *performances* across sections. Finding substantial reductions in inter-section variability would call attention to those portfolios in which inter-section differences drop; what are the reasons? Perhaps, in certain portfolios, lack of evidence about the aspect of skill targeted in a given section can be mitigated by evidence in a different section. If so, the added assurance of the validity of ratings may justify the logistical difficulties and additional resources necessary to bring the most highly variable portfolios back for a second, cross-sectional, look.

The FACETS program produces indices of the consistency of agreement across readers and portfolio sections for each student, analogous to the reader fit indices previously discussed. Again, we focus on the information-weighted mean-square fit statistic, or "infit." Again, the expectation is 1.0, the range is 0 to infinity, and the higher the value, the more the variability in the rating pattern even when section and reader effects are taken into account. When sections and readers are fairly similar, an infit mean-square less than 1 indicates little variation in the ratings (i.e., a "flat-line" profile consisting of all or nearly all the same rating), while an infit mean-square greater than 1 indicates more than typical variation in the ratings (i.e., a set of ratings with one or more unexpected or surprising ratings, ratings that don't seem to fit with the others).

And again, there are no hard and fast rules for upper and lower control limits for the student infit index. Some programs in our experience have used an upper control limit of 2 and a lower control limit of .5; more stringent limits

might be instituted if the goal is to strive to reduce significantly the variability within a system. As with reader fit indices, the more extreme the value, the greater the potential gains for improving the system—either locally, by rectifying an anomaly with a second look at an individual student's work, or globally, by gaining insights to improve training, rating, or logistic procedures.

We adopted an upper control limit for the student infit mean-square of 2.0 in this study, a liberal upper control limit to accommodate some variability in the 13 ratings given to an individual student. We wished to allow for a certain amount of variation in readers' perspectives across the ratings, yet still catch cases in which disagreement was "out of statistical control" (Deming, 1975). An infit value beyond the upper control limit signals a portfolio that might call for a second look before the student's final score is issued, particularly if it is near a cut point.

Of 3889 portfolios, 225, or 6%, had infit values at or above 2.0 before within-section discrepancies were resolved. In 85 of them, the lack of fit was caused by subsequently resolved discrepancies *within* sections. In the other 140, the cause was variability *across* sections, with ratings within each section agreeing within a point.[11] Given this relatively small number of portfolios, it would seem feasible to ask readers to review all of them, or at least those near cut-points, to see if the variability in the rating profiles can be explained on the basis of the work. Those profiles that cannot be explained are *inter*-section discrepancies. Like the *intra*-section discrepancies flagged in the current system, they merit the Chief Reader's attention—to initiate action perhaps at the individual level, by moderating the ratings for this student, or perhaps at the system level, by sparking discussions to clarify standards or improve reader training.

The rating data shown in Table 11 include all four combinations of whether or not a pattern exhibits a discrepancy, and whether it has a high or low fit index. The ratings of Student #682 are most typical; no discrepancy and a moderate fit index (1.0). There is some variation in ratings: eight of the 13 ratings are 2's, four are 1's, and one is a 3. In each section, though, all the ratings are the same or within one point.

Student #852, has a similar mix of rating values—seven ratings are 2's, five are 3's, and one is a 1—but now a "discrepancy" exists because Section A has both a 1 and a 3. As it happens, the 3 was given by Reader 12, the second most lenient in the group; the Chief Reader eventually replaced it with a 2. Reducing the discrepant rating and treating the revised number at face value has, in this case, the same effect as concluding the rating is not unusual in light of the reader's leniency and providing a model-based final score that takes the leniency into account. The end is the same, although in an automated system the model-based action would have consumed fewer scarce resources, namely, the attention of the Chief Reader. The moderate fit index indicates that this

**Table 11**

**Ratings Patterns and Fit Indices of Selected Students**

| | Section A | | | Section B | | Section C | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Student #682 (Infit=1.0, Discrepancy=No; $\hat{\theta}$=.11, SE=.46) | | | | | | | | | | | | |
| Reader Harshness Measure | .18 | .00 | .29 | -.26 | -.77 | .18 | -.09 | .18 | -.09 | .18 | -.09 | .18 | -.09 |
| Reader ID | 13 | 10 | 80 | 11 | 17 | 13 | 82 | 13 | 82 | 13 | 82 | 13 | 82 |
| Ratings | 1 | 2 | 2 | 3 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Student #852 (Infit=1.0, Discrepancy=Yes; $\hat{\theta}$=1.01, SE=.44) | | | | | | | | | | | | |
| Reader Harshness Measure | .00 | .12 | -.66 | .19 | .29 | .00 | .52 | .00 | .52 | .00 | .52 | .00 | .52 |
| Reader ID | 10 | 18 | 12 | 26 | 80 | 10 | 20 | 10 | 20 | 10 | 20 | 10 | 20 |
| Ratings | 2 | 1 | 3 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 3 | 3 | 3 |
| Student #3751 (Infit=4.5, Discrepancy=No; $\hat{\theta}$=.49, SE=.45) | | | | | | | | | | | | |
| Reader Harshness Measure | .30 | .34 | -.09 | -.33 | -.21 | .18 | -.26 | .18 | -.26 | .18 | -.26 | .18 | -.26 |
| Reader ID | 16 | 14 | 82 | 19 | 28 | 13 | 11 | 13 | 11 | 13 | 11 | 13 | 11 |
| Ratings | 1 | 1 | 1 | 4 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| Student #1377 (Infit=2.3, Discrepancy=Yes; $\hat{\theta}$=.30, SE=.47) | | | | | | | | | | | | |
| Reader Harshness Measure | .34 | -.25 | -.03 | .52 | -.11 | .38 | .34 | .38 | .34 | .38 | .34 | .38 | .34 |
| Reader ID | 14 | 22 | 81 | 20 | 29 | 15 | 14 | 15 | 14 | 15 | 14 | 15 | 14 |
| Ratings | 3 | 4 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |

pattern is only anomalous at the surface level of raw ratings; since it is *not* anomalous in terms of the main-effects model. If desired, such patterns can be dealt with automatically through use of the model-based rather than observed student scores, releasing the Chief Reader's attention for more challenging cases (e.g., a 1 and a 3 that cannot be explained by main effects of readers).

Student #3751 had the highest infit mean-square value in the data set, 4.5, even though ratings were consistent within sections. The two 4 ratings in Section B are quite unexpected in light of the 1's in all the other sections. The two readers who gave the 4's were among the more lenient, but even so, the FACETS table of atypical ratings showed model expectations of 2's, not 4's, for these combinations of student, reader, and section estimates. Did the student truly excel in Section B in comparison to the other sections, or do the 4's reflect idiosyncratic rating behavior on the part of the two readers who rated Section B? The former seems more likely, but we would need to look at the portfolio to make the critical determination.

We would also want to reexamine the work of Student #1377, who had a fit value of 2.3. This time there is a discrepancy within Section A (3, 4, 2) as well as an unevenness across profiles (mostly 1's everywhere else).[12] The Chief Reader resolved the discrepancy by changing the 4 to a 3 (the reader who originally gave it was slightly above average in harshness), but again this process does not address the unevenness across sections.

In addition to flagging unexpected rating patterns, fit indices can also be used to condition the degree of confidence we place in an overall score—more confidence for internally consistent patterns, less for variable ones. An approximate adjustment is obtained by multiplying the standard error of measurement associated with a given estimate by the square root of the fit mean-square value. In the examples above, there would be no modification of the standard errors of Students 852 and 682, but inflation factors of 2.1 and 1.5 for those of Students 3751 and 1377, respectively.

**Discussion**

This section offers comments of three types. The first type concerns ways that a model-based analysis such as the one described above can be used to fine-tune AP Studio Art in its present form. The last speculates on possibilities that would arise in on-line, real-time rating environments, freed of the logistical constraints of the current system. In between, we address a more focused question that arises in either case, namely, how to deal with discrepancies.

*Comments apropos the current system*

Sections are rated separately and independently in the present system. All Section A submissions, for example, are laid out in a huge gymnasium. After three successive readers rate the works in each row, aides collate the ratings and

bring discrepancies to the Chief Reader for resolution. This arrangement facilitates standard-setting sessions (because all readers can discuss a section together before they rate it) and the handling of within-section discrepancies (because they can be brought to the Chief Reader's attention quickly). It is difficult, however, to integrate information across sections or to evaluate readers' patterns across all the portfolios they rate. It would be possible under the present system to analyze each section's ratings at the end of its session (perhaps overnight), for each section by itself and for each section along with those that had been rated previously. The results of these analyses could serve the following purposes:

- Portfolios that did not appear as discrepancies yet merited the Chief Reader's attention could be flagged. This would include particular portfolio sections that did not exhibit discrepancies, but were rated by extreme combinations of harsh or lenient readers, as well as portfolios with strikingly uneven profiles up to that point in the reading.

- Summary reports that condensed information from *all* ratings by each reader, in terms of harshness and fit, would be available to the Chief Reader to supplement impressions gleaned from discrepancies. The readers themselves could receive these reports to help them monitor their own work (as described in Stahl & Lunz, 1991).

- After the last section is read, final reported scores could be based on model-based measures (weighted as desired) rather than observed scores, in order to take remaining reader-harshness main effects into account.

- Through the use of common readers and/or portfolios over assessment years, a common scale over time could be established statistically as well as through standard-setting sessions. Rather than having to determine cut points from each year's weighted score independently of previous years' data, cut points could be pre-established on the $\theta$ scale. This would facilitate comparisons of performance over time, and make it possible to know before all the ratings were in whether a given student were near a cut point.

*Should discrepancies be resolved?*

One of the key ways to improve a system is to identify and investigate unusual occurrences. Local problems may be detected and corrected; more importantly, clues for improving the system are often manifest. The within-section discrepancy procedure serves this crucial role in the present system. The Chief Reader is in this way provided incidents that, compared to a rating chosen at random, are more likely to involve a performance that is "hard to rate" or a reader who is out of synch with the others. As it has evolved, the current definition of "a discrepancy" calls a manageable number of portfolios to the Chief Reader's attention. The value of this monitoring and fact-finding function of the discrepancy procedure lies beyond question (although the

identification of anomalies could be fine-tuned with the model-based procedures discussed above).

But this system-level function is distinct from the local function of verifying the quality of the ratings of the individual portfolios involved in the discrepancies. The foregoing analyses suggests it may not be necessary, nor perhaps even desirable, to have the Chief Reader "resolve" all discrepancies (that is, to replace one or more ratings with his or her own). It is difficult to avoid the implication that the "resolved" ratings were in some way "wrong." But some variation among readers is expected, and is addressed largely by soliciting ratings from several readers. Given the usual amount of variation and the existing (modest) variation in reader harshness, discrepancies will occur at predictable rates and not all of them are problematic (e.g., Student #852). By evaluating a profile of ratings in light of all sections and individual readers, against the context of typical variation, the Chief Reader would have a more extensive foundation for determining whether a given discrepancy is indeed the occasional aberration. Only the most extreme instances would then warrant excision and replacement. A milder remedial action would be for the Chief Reader to add his or her rating to the profile, reasoning that work that evokes a broader spread of ratings requires a larger sample of readers to maintain quality.[13] The remaining discrepancies could be allowed to stand as is, say, when the fit indices for the profile as a whole fell within control limits.

*Comments apropos an on-line, real-time system*

In an "on-line, real-time system," performances and readers can be brought together in any combination at any time during the process, and ongoing analyses can examine ratings as they accumulate to guide subsequent rating assignments. Elements of such a system are currently being piloted with computer-image readings of National Assessment writing responses (Johnson, 1993). In that project, images of students' essays are scanned into a computerized database, and can appear on the terminals of readers to whom they are assigned without the need for further paper-handling. (Technology currently exists to capture spoken or written responses directly from students, without intervening paper or audiotape.) The readers enter their ratings into the database, which can be analyzed at any point in time. In time, it may be feasible to work with digitized versions of AP Studio Art portfolios, to provide any reader access to any section of any portfolio at any point in time. Whether or not this scenario lies in the future of AP Studio Art, we can contemplate test theory and quality control issues in such a system.

An on-line, real-time system opens the door to dynamic allocation of the most scarce resource in the system, namely, the expert readers. The statistical foundations of dynamic sampling extend back to the Sequential Probability Ratio Test that Abraham Wald developed during World War II to increase the efficiency of testing torpedoes (since the testing process itself was destructive) (Wald, 1947).

The key idea is analyzing data as they are collected, to guide which observations to make next and when to stop. Some cases will require more data and others less, but data-gathering resources will have been allocated to where the need is greatest. Impressive efficiencies can be obtained, compared to gathering the same amount of the same kind of data for all cases.

In AP Studio Art, as an example, we might begin by soliciting one rating for each section and then check estimated Student Measures and indices of fit. This small number of ratings might provide sufficient evidence for the final score of a very high or very low portfolio, if the ratings were consistent across sections. The criterion would be whether the present estimate fell within a scoring category with sufficiently high probability, as gauged by its estimation error (say, within 1.5 fit-adjusted standard errors from the reported-score cutoffs). Additional readers would be assigned to most portfolios, across sections according to the information they'd provide—a function of the level of performance and the consistency of ratings across and within sections. Portfolios with uneven profiles or discrepant ratings within sections would require more readings, especially when the projected weighted scores fell near cut points. The degree of decision consistency of the current system could be attained with fewer readings per portfolio on the average, although some hard-to-rate portfolios could end up with more readings than in the current system.[14]

An on-line system would also facilitate another option (which could be employed somewhat less conveniently within the current system): Readers could themselves flag portfolios to which they have been assigned but that they feel either discomfort in rating or a lack of expertise about the styles or media represented in the work. A more appropriate reader would then be assigned. The test-theory assumption of interchangeable readers is convenient for quality control, and we do in fact want readers to rate pieces within the same framework of meaning—but readers are obviously and inevitably unique, often in ways they are quite aware of, in ways they could exploit to improve the validity of the scores. (In such cases, analyses with respect to subsets of informed readers would replace analyses with respect to the group at large to help monitor quality.)

## Conclusion

The attractive features of performance assessment include the potential for instructional value and the elicitation of direct evidence about constructive aspects of knowledge. Principal concerns include weight of evidence and accountability. The approach described above addresses aspects of both of these concerns, for it is only by working back and forth between such statistical and naturalistic analyses that a common framework of meaning can be established, monitored, and improved over time. We have stressed the importance of integrating statistical quality control information and substantive knowledge of the system because, in our view, either type of knowledge by itself is insufficient. Naturalistic and statistical approaches to studying complex phenomena must be viewed as complementary, not competing approaches. By employing both types of approaches to study a performance assessment system, one can draw on the strengths of each approach to examine the system from a number of different angles.

Resources for assessment will always be limited, and performance assessments demand more resources than multiple-choice assessments. Those of us working with performance assessment have a responsibility to consider carefully how to make best use of those limited resources. That means knowing what aspects of the performance we want to make inferences about, so that we don't waste resources gathering data that holds little value as evidence for our purposes. To guide our actions, we need an overarching conceptual framework for collecting, analyzing, and integrating various types of information about a performance assessment system. We must identify the kinds of information we need to collect about our system to help us determine whether the system is functioning as intended. We must learn to use this information to improve the system. We have attempted to lay out a framework that addresses these needs.

In this presentation, we have shared findings from an analysis of rating data from the AP Studio Art portfolio assessment. Using the FACETS program, we have illustrated an approach for establishing a quality-control overview for a rating procedure and monitoring its effectiveness. The emphasis is on examining the role of each facet in a rating procedure in order to gain a better understanding of how it operates and where it fails. The output from the analysis explicates current operation of the process and provides clues that can help us improve the accuracy, accountability, and fairness of our assessment system.

A statistical framework such as the one provided by FACETS in this study can highlight observations within a complex system that lie outside the usual ranges of variability—those that are unexpected, unusual, or surprising. Once we have identified unexpected occurrences, we need to determine why they occurred. That requires a different approach to data gathering. We need to go beyond the rating data to ask the important "why" questions of those involved in the performance assessment—the readers and the students. The selections from our analysis of interview protocols illustrate the kinds of insights one might gain by engaging readers in discussions about hard-to-rate portfolios. Through

these discussions we are discovering why certain portfolios are more difficult to rate than others. We are gaining a better understanding of the cognitive processes that underlie the unusual or discrepant rating patterns we observed in the data. We are learning about similarities and differences in individual readers' views of the works, their understandings of the meaning of the three sections to be evaluated, and their uses of the various rating scale points. In short, through our analysis of the interview protocols we are obtaining much valuable feedback about how the rating procedure is operating from the perspectives of the readers. This feedback can help sharpen the definitions of the various rating dimensions, refine scoring rubrics, and improve reader training.

When scores are based on human judgments, validation requires that the ways of gathering and evaluating evidence follow regular and verifiable procedures, and that these procedures satisfy quantifiable and acceptable standards at the level of the system. We believe that the framework illustrated here represents a sensible and fiscally responsible approach for studying complex performance assessment systems. By making public the materials and results of an analysis such as this, one can demonstrate (1) the nature of the evidence the assessment evokes, (2) the processes by which statements about students' performances are obtained, (3) the degree of accuracy of these statements, and (4) the procedures for identifying and rectifying irregularities. "Statistical analysis versus reflection and debate" is a false dichotomy. Constructing a statistical framework in this manner does not supplant, but rather supports and evidences, the entry and discussion of questions of value and standards of good work.

# References


Askin, W. (1985). *Evaluating the advanced placement portfolio in studio art.* Princeton, NJ: Educational Testing Service.

Bleistein, C. A., Flesher, R. B., & Maneckshana, B. (1993). *Test analysis, College Board Advanced Placement Examination: Studio Art General/Studio Art Drawing* (Report No. SR-93-82). Princeton, NJ: Educational Testing Service.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*(1), 1-18.

Carnes, V. (1992). *Teacher's guide to advanced placement courses in studio art.* New York: College Entrance Examination Board.

College Entrance Examination Board. (1993). *1993-94 Advanced placement studio art* [poster]. New York: Author.

College Entrance Examination Board. (1994). *Advanced placement course description: Art.* New York: Author.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, H. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Deming, W. E. (1975). On some statistical aids toward economic production. *Interfaces, 5,* 1-15.

Deming, W. E. (1980). *Scientific methods in administration and management.* (Course No. 617). Washington, DC: George Washington University.

Englehard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Fisher, R. A. (1973). *Statistical methods and scientific inference.* New York: Hafner.

Houston, W. M., Raymond, M. R., & Svec, J. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement, 15*(4), 409-421.

Johnson, E. G. (1993). *The results of the NAEP 1993 field test for the 1994 National Assessment of Educational Progress.* Princeton, NJ: Educational Testing Service.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5-16.


81

LeMahieu, P., Gitomer, D., & Eresh, J. (in press). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice.*

Linacre, J. M. (1989). *FACETS* [Computer program]. Chicago, IL: MESA Press.

Linacre, J. M. (1993). *Many-faceted Rasch measurement.* Chicago, IL: MESA Press.

Linacre, J. M., & Wright, B. D. (1994). *A user's guide to Facets: Rasch measurement computer program* [Computer program manual]. Chicago, IL: MESA Press.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*(4), 425-444.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3,* 331-345.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools.* New York: The Free Press.

Mitchell, R., & Stempel, A. (1991). *Six case-studies of performance assessment.* Washington, DC: Office of Technology Assessment.

Resnick, L. B., & Resnick, D. P. (1989). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Conner (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer.

Stahl, J. A., & Lunz, M. E. (1991, April). *Judge performance reports: Media and message.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Educational Research* (Vol. 17, pp. 31-74). Washington, DC: American Educational Research Association.

# Appendix A

Letter of Introduction
Description of the Reseach Study
AP Studio Art Reader Survey

EDUCATIONAL TESTING SERVICE PRINCETON, N.J. 08541

609-921-9000
CABLE-EDUCTESTSVC

June 3, 1992

Dear AP Studio Art Readers,

Welcome to the 1992 AP Studio Art Reading. We're glad you're here, and we look forward to a productive time together.

In this packet are several documents that we'd like you to read upon arrival. First, you will find a description of a research study that two ETS researchers, Carol Myford and Bob Mislevy, will be carrying out during the AP Studio Art reading. We are very enthusiastic about this research and feel that it will provide us with much useful information about our assessment process. After reading the description of the study, you may have some questions or concerns. If so, please do not hesitate to raise them with the researchers or with us. We want you to feel sufficiently informed about and comfortable with the research activities that are planned.

Second, you will find in this packet an AP STUDIO ART READER SURVEY. We would request that you take 15-20 minutes to complete this survey and turn it in by 4:00 p.m. on Thursday, June 4. ETS is collecting this information to meet several needs. The researchers will use some pieces of the information; AP Studio Art program staff will use other pieces for program planning and reader recruitment/selection purposes. We ask your cooperation in helping us to gather this important information.

Thanks in advance for all your hard work and dedication. We really appreciate your efforts!

Sincerely,

Alice Sims-Gunzenhauser

Ray Wilkins

# ISSUES AND TECHNICAL PROCEDURES IN RATING PORTFOLIOS

Carol M. Myford and Robert J. Mislevy

As alternative assessment forms such as ratings of extended performances and portfolios become increasingly attractive, issues involved with defining rating standards and assuring consistency in their use over tasks and raters become increasingly important. The proposed project will explore issues in the context of AP Studio Art from two complementary perspectives: statistical and naturalistic.

We want to develop a framework for monitoring and continually improving performance rating systems. This requires integrating statistically information and substantive knowledge--in this case, expertise about the standards of AP Studio Art. Either type of knowledge by itself is insufficient. The "statistical" aspect of the project will take place later, analyzing data from the June '92 reading to address generalizability issues and to develop analytic tools. The "naturalistic" aspect begins now, during the reading period itself.

The best way to develop the meaning of rating scale values is through discussions and examples, to promote among judges a shared view of what to look at and what is important, and a common language for their evaluation. Our idea is to engage readers in discussions about the qualities and distinctions they use in making their judgments. We will use hard-to-rate portfolios to stimulate these discussions. Portfolios that are atypical in appearance or that have provoked different points of view push the boundaries of scoring procedures and reveal the most about the discriminations and the summaries judges must make to arrive at their final ratings.

We hope these discussions will provide insights into the real processes that underlie the patterns in ratings statistics, to illuminate our understanding of the various dimensions that are to be assessed and of how scale points are operationally defined. The results should prove useful not only to AP Studio Art, but much more generally for any programs that use dimension definitions, scoring rubrics, and expert judgments.

Procedure:

1. Identify 9 portfolios each in Section A and Section B during the June 1992 AP Studio Art general portfolio reading that evoke "interesting differences" among judges. For example, some portfolios show 2-point or even 3-point discrepancies in the ratings judges give on one or more dimensions. Among portfolios showing 1-point discrepancies, a more common occurrence, some are highly individualistic in style, and may stimulate fruitful discussions. We will make sure that at least some of the portfolios we select have been resolved as 2's. By focusing on such portfolios, we may be able to identify features of borderline "high 2's" and "low 2's" that might be used to develop a more finely graded scoring rubric.

2. For each portfolio, meet with the two judges who rated the portfolio and discuss the rationales for their judgments. Probe for similarities and differences in their views of the works, their understandings of the meaning of the various rating dimensions, and their uses of the various rating scale points. Tape record the discussions.

3. Review the tapes of the discussions in order to gain insights into the commonalties and the differences among judges as to how they view the rating dimensions and how they employ the rating scale points.

4. Prepare an informal interim report which summarizes what was learned in these discussions to share with the AP Studio Art Chief Reader, Table Leaders, and ETS program staff.

5. Write a formal report for general dissemination, integrating the naturalistic and statistical perspectives.

# AP STUDIO ART READER SURVEY

June 1992

*Instructions: Please provide all the information requested. If you have any questions, Alice Sims-Gunzenhauser and Ray Wilkins are available to help. When you have completed the survey, return it to Alice or Ray by 4:00 p.m. on Thursday, June 4.*

Name: _____ Reader ID: _____

1. ACADEMIC DEGREES (Certificates or relevant credentials)

| Degree | Year | Institution | Major | Minor |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |

2. TEACHING EXPERIENCE (at the collegiate and/or secondary school level)

| Dates | | Institution |
|---|---|---|
| From | To | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

3. In the first column of the table below you will find a number of art disciplines listed. Indicate which disciplines you teach or have taught by checking ( ) those disciplines in Column 2. Indicate the areas in which you personally work as an artist by checking those disciplines in Column 3.

| | I have taught courses in the following area(s): | My own work is in the following area(s): |
|---|---|---|
| | | |
| Drawing | | |
| Painting | | |
| Sculpture | | |
| Printmaking | | |
| Design | | |
| Photography | | |
| Video | | |
| Filmmaking | | |
| Metal | | |
| Wood | | |
| Clay | | |
| Glass | | |
| Ceramics | | |
| Fiber Art | | |
| Computer art | | |
| Conceptual art | | |
| Other (please specify) : | | |
| | | |

4. List below any short-term, art-related workshops or seminars you **have taught** in the last two years.

5. List below any short-term, art-related workshops or seminars you **have taken** in the last two years.

6. Do you currently teach a course that is equivalent to AP Studio Art in a college, university, or art school? *(Check one.)*

_____ (1) Yes
_____ (2) No *(If no, skip to question 7.)*

6a. Does your college/university/art school give credit for AP Studio Art? *(Check one.)*

_____ (1) Yes
_____ (2) No
_____ (3) Don't know

6b. If yes, what grade must a student receive in order for credit or advanced placement to be awarded? *(Check one.)*

_____ (5) 5
_____ (4) 4
_____ (3) 3
_____ (2) Don't know

7. Have you given AP Studio Art workshops or institutes? *(Check one.)*

_____ (1) Yes
_____ (2) No

8. If you teach at the secondary level, how many years have you taught AP Studio Art? *(Check one.)*

_____ (1) Less than 1 year
_____ (2) More than 1 year, but less than 3 years
_____ (3) More than 3 years, but less than 5 years
_____ (4) More than 5 years, but less than 10 years
_____ (5) More than 10 years
_____ (6) I do not teach at the secondary level.

9. How many years have you served as an AP Studio Art reader? *(Check one.)*

_____ (1) Less than 1 year
_____ (2) More than 1 year, but less than 3 years
_____ (3) More than 3 years, but less than 5 years
_____ (4) More than 5 years, but less than 10 years
_____ (5) More than 10 years

10. Have you shown your art work within the last five years? *(Check one.)*

_____ (1) Yes
_____ (2) No *(If no, skip to question 11.)*

10a. If yes, where have you exhibited?

10b. Have you recently received any awards or prizes?

_____ (1) Yes
_____ (2) No *(If no, don't worry! Neither have we!)*

10c. If yes, list those awards and/or prizes.

11. Provide a brief description of the type of art work you do.

12. What is your primary medium?

13. Are there other media that you sometimes use?

_____ (1) Yes
_____ (2) No *(If no, skip to question 14.)*

13a. If yes, what other media do you use?

14. Which artists do you most admire?

15. What are the primary influences on your work, if any?

15. Give a short description of your philosophy of art or your intentions as an artist.

***Thanks for taking the time to respond to this survey. We really appreciate your efforts!***

91

# Appendix B

Discussion Leaders' Protocol

ISSUES AND TECHNICAL PROCEDURES IN RATING PORTFOLIOS

Discussion Leaders' Protocol

[Note: These activities are viewed as discussions rather than interviews. Rather than following a strict interview protocol, we want to stimulate a discussion that evokes readers' comments about how they view the rating scales, sparked by portfolios that have proven to be "hard to rate." A hand-out has been prepared for the readers that will help them enter the discussion with this perspective. The questions below are intended to help the discussion get started and shape its general direction and to make sure that particular points are addressed for particular types of portfolios that will have been selected.]

Opening statement:

**I'd like to start by thanking you for helping with this project. AP Studio Art is getting a lot of attention these days. It goes far beyond multiple-choice questions. It integrates assessment with learning; students produce valued pieces of art in the process. What's most important for our project is that scores must be based on the judgments of experts such as yourselves. We need to learn more about the process of defining and using rating scales, not just for possible improvements for AP Studio Art, but as a prototype for the same kinds of challenges we'll confront in new kinds of assessments in areas like writing, science, and mathematics.**

**We've asked you to come to talk with us so we can learn more about how you view and use the Section (A/B) rating scale. To help do this, we've picked one of the portfolios you read that turned out to be "hard to rate;" that is, it provoked differences in ratings among different readers.**

**Can you describe what is running through your mind as you look at this portfolio?**

**The working draft of a scoring rubric lists five distinct areas to look at when you make ratings. Are these useful ways of thinking about this particular work? If so, did any of these aspects really jump out at you when you judged this portfolio? If not, what did capture your attention as you looked at this portfolio?**

Optional prompts, depending on where things go next:

**Did you feel that work was strong in some of these aspects [of the rubric] but weak in others?**

If "yes,"

**Which did you feel were more important when it came to giving a final rating?**

**Would you call this a "typical" portfolio or an "unusual" one?**

If "unusual,"

**In what ways?**

**How do you approach portfolios that aren't like most others? Are there certain criteria you use when evaluating "unusual" portfolios that you don't use when evaluating "typical" portfolios?**

**As you think about your own training, experience, and teaching style, are there aspects of your background that lead you to view this portfolio in a special way? For example, is it in a style or a medium that's particularly familiar or unfamiliar to you, or do you have strong feelings about it for some other reasons?**

For portfolios resolved as 2's:

**In the past, nearly half of the submissions ended up with scores of 2--as did this one. Given that it was one that provoked differences of opinion, do you think you might call it a "high 2" or a "low 2?"**

**What qualities might you point to that might be used in making high2/low 2 distinctions among other portfolios?**

For Section B portfolios:

**In Section B, the students provide short descriptions of what they were trying to accomplish. Did this play a role in your thinking?**

If "yes,"

**In what ways?**

**In general, how much of a role do the paragraphs usually play in your judgments? Are there situations in which you pay particular attention to them?**

**The draft of the rubric is written generally to apply to all portfolio sections. What is especially important to look at in Section B? Is the general rubric applicable to Section B? Are there aspects of Section B judging that aren't well captured in this general rubric that might be added to apply specifically to Section B?**

# Endnotes

[1] An alternative approach often used to model rating data is Cronbach, Gleser, Nanda, & Rajaratnam's (1972) generalizability theory—G-theory, for short. In G-theory, the dependent variable is the rating in its original ordinal units, modeled in terms of additive fixed and/or random effects for aspects of the rating occasions. In FACETS, the dependent variable is log-odds of conditional probabilities of adjacent rating categories, modeled in terms of additive fixed effects. Compared to the present study, typical applications using G-theory place greater emphasis on modeling and estimating variance components, for the purposes of explicating sources of uncertainty and planning subsequent studies in light of existing patterns and magnitudes of variation. Relatively less emphasis is placed on estimating effects and identifying atypical patterns associated with particular readers, tasks, reader/portfolio combinations, etc., for the purpose of identifying leverage points for modifying the system.

[2] In 1993, the AP Studio Art Development Committee made the decision to combine the "Color" and "Design" subsections of Section C into a single subsection called Color/Design. Students now submit slides of eight works in which color and design principles are the focal point.

[3] Since Color and Design were collapsed into a single subsection in 1993, each of the two readers of Section C of a portfolio now gives three ratings, so that Section C receives 6 ratings altogether and the portfolio as a whole receives 11.

[4] Analogous discussions in an operational system would involve the original readers, as one aspect of an integrated system growing from the coordinated efforts of readers, program administrators, and statisticians.

[5] The arts educators who read the AP Studio Art portfolios have had a good deal of experience in critiquing works of art prior to their serving as readers. Classroom critiquing sessions provide a natural training ground for students of the visual arts to learn to make aesthetic judgments about works of art. Students learn to be tolerant of other's judgments that may differ from their own. They come to realize that there can be multiple legitimate responses to a work of art, and that each is to be respected in its own right. These are abilities that are highly valued in the visual arts community. They are marks of professionalism in this field. In the context of AP Studio Art, readers know that when a portfolio receives discrepant ratings, they may be called upon to "talk through" their different views of the works, describing their rationales for the ratings they gave as the Chief Reader seeks to find common ground. These are understood to be the "rules of the game" in the AP Studio Art program. People who don't have this attitude rarely seek to become AP Studio Art readers, and if they do, they generally don't remain so for long.

[6] Although even more time than this *is* spent altogether in the discussions the students have with their teachers throughout the year, as they create their works! The goal would be for these broader instructional discussions to be consonant with the qualities and standards reflected in the readers' discussions.

[7] Analyzing ratings in their original ordered scales presumes the rating categories are "equal interval" scores. FACETS attempts to fit a model with unobservable variables that are equal-interval measures with respect to logits of conditional probabilities, with rating-category parameters that moderate these measures' relationships to observed ratings.

[8] FACETS calculates several fit statistics. "Infit" and "outfit" are estimates of residual mean squares and are computed as chi-square statistics divided by their degrees of freedom; infit weights residuals by their modeled variances, outfit doesn't. Standardized versions of both are also reported. Outfit tends to be more sensitive to extreme outliers, but with the range of parameters and ratings in our data, infit and outfit were practically indistinguishable. For simplicity, we discuss only infit, only in the original mean-square version.

[9] This figure was created by generating, for each student, two independent normal values with the mean equal to the student's estimated $\theta$ and standard deviation equal to the standard error.

[10] Gauging effects in terms of variance ratios, as in G-theory, we find that adjusting for reader effects would not materially improve the accuracy of scores for this program. A variance ratio analogous to reliability *without* adjustment for reader effects can be calculated as follows:

$$\rho = \frac{\text{true-score variance}}{\text{true-score variance} + \text{SEM}^2 + (\text{reader variance})/7} \approx \frac{1.36^2}{1.36^2 + .47^2 + .31^2/7} = .887.$$

*With* adjustment for reader effects, the reader variance term disappears, yielding a value of .893.

[11] The 140:85 ratio of inter- and intra-section discrepancies accounting for high fit indices is consistent with the higher correlation of readers for the same section than for different sections. Separate FACETS runs for each section would, of course, flag only intra-section discrepancies, taking reader main-effects into account. These would generally be the same portfolios that are currently flagged, in terms of observed-rating discrepancies, with two occasional differences: (1) a one-point observed difference could have a high fit value if the low rating was from a very lenient reader and the high rating was from a very harsh one; and (2) a two-point discrepancy could have a low fit value if the low rating was from a harsh reader and the high rating was from a lenient one.

[12] Note that the discrepancies for Students 852 and 1377 both occurred in Section A. Even though inter-rater correlations are about the same for all sections, one can anticipate more discrepancies in Section A than in other sections simply because more ratings are given—three, as opposed to two. Using a model-based approach to flag outliers rather than the present observed-discrepancy approach would more greatly reduce calls concerning variation within Section A than calls concerning variance within the other sections.

[13] The additional rating could be incorporated into either the current weighted score or a model-based student-measure estimate, by assigning weights that maintained the current section totals but spread across more ratings.

[14] Again, variable weightings for individual ratings could be used to maintain the targeted relative weight of information from the various sections.